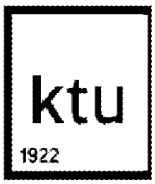


KAUNO TECHNOLOGIJOS UNIVERSITETAS
INFORMATIKOS FAKULTETAS

Ernesta Kebelytė

LIETUVIŠKO AUTOMATINIO NAUJIENŲ AGREGATORIAUS PROTOTIPAS

Baigiamasis magistro darbas

Vadovas
dr. Mantas Lukoševičius

KAUNAS, 2015

KAUNO TECHNOLOGIJOS UNIVERSITETAS
INFORMATIKOS FAKULTETAS

LIETUVIŠKO AUTOMATINIO NAUJIENŲ AGREGATORIAUS PROTOTIPAS

Baigiamasis magistro darbas
Programų sistemų inžinerija (621E16001)

Vadovas
(parašas) dr. Mantas Lukoševičius
(data)

Recenzentas
(parašas) dr. Lina Bisikirskienė
(data)

Projektą atliko
(parašas) Ernesta Kebelytė
(data)

KAUNAS, 2015

KAUNO TECHNOLOGIJOS UNIVERSITETAS

Informatikos fakultetas

(Fakultetas)

Ernesta Kebelytė

(Studento vardas, pavardė)

Programų sistemų inžinerija (621E16001)

(Studijų programos pavadinimas, kodas)

Baigiamojo projekto „Lietuviško automatinio naujienų agregatoriaus prototipas“

AKADEMINIO SĄŽININGUMO DEKLARACIJA

20 15 m. gegužės 25 d.

Kaunas

Patvirtinu, kad mano, **Ernestos Kebelytės**, baigiamasis projektas tema „Lietuviško automatinio naujienų agregatoriaus prototipas“ yra parašytas visiškai savarankiškai ir visi pateikti duomenys ar tyrimų rezultatai yra teisingi ir gauti sąžiningai. Šiame darbe nei viena dalis nėra plagijuota nuo jokių spausdintinių ar internetinių šaltinių, visos kitų šaltinių tiesioginės ir netiesioginės citatos nurodytos literatūros nuorodose. Įstatymų nenumatytų piniginių sumų už šį darbą niekam nesu mokėjęs.

Aš suprantu, kad išaiškėjus nesąžiningumo faktui, man bus taikomos nuobaudos, remiantis Kauno technologijos universitete galiojančia tvarka.

(vardą ir pavardę įrašyti ranka) *(parašas)*

Turinys

Lentelių sąrašas

Paveikslėlių sąrašas

TERMINŲ IR SANTRUMPŲ ŽODYNAS

HTML	Programavimo kalba skirta internetinių puslapių kūrimui (angl. Hyper Text Markup Language).
Kamienizavimas	Žodžio galūnės šalinimas taip įgalinamas tą pačią reikšmę, bet skirtingas galūnes turinčių žodžių sujungimas.
Personalizavimas	Kompiuterinės sistemos gebėjimas prisitaikyti prie skirtingų vartotojų nustatymų.
Prasminiai žodžiai	Žodžiai, kurie naujienoje pasikartoja daugiau nei vieną kartą neįkaitant nereikšminių žodžių.
Raktažodis	Bet kuris žodis pasitaikęs naujienų tekstuose neįskaitant nereikšminių žodžių.
RSS	Informacijos, pateikiamos „XML“ formatu, pateikimo iš internetinių portalų technologija.
Naujienų grupė/kategorija	pasitelkiant tekstinių dokumentų grupavimo algoritmus gauti skirtingi dinaminei naujienų klasteriai.
Naujienų agregatorius	Internetinė paslauga arba paprasčiau svetainė, kuri surenka informacijai š daugybės šaltinių ir parodo ją viename lange.
Naujienų klasifikavimas	Naujienų tekstų grupavimas į nustatytas klases (pvz. „Sportas“, „Verslas“).
Nereikšminiai žodžiai	Žodžiai, kurie tekstuose pasikartoja dažniausiai, tai įvairūs jungtukai, įvardžiai ir kita.
XML	Bendros paskirties duomenų struktūrų bei jų turinio aprašomoji kalba.

Kebelytė, E. Lietuviško automatinio naujienų agregatoriaus prototipas. Programų sistemų inžinerijos magistro baigiamasis projektas / vadovas dr. Mantas Lukoševičius; Kauno technologijos universitetas, Informatikos fakultetas.

Kaunas, 2015. 62 p.

SUMMARY

Everyday there is a huge amount of information being created, which would be impossible to cover without any software tools. As one of those tools helping to assemble and display this information there is RSS (Rich Site Summary) subscriptions. This technology is designed for fast and secure assembly and display of newest news entries from different news sources. The aim of the system - group Lithuanian news from different news sites, so that the users could effectively and faster find news unique to them. The information is grouped into categories, but leaves a lot of space for interpretations: what the groups should be, how to join or disjoin different information.

In order to create smart news aggregator first analysis of grouping methods, algorithms tools was performed. According to this analysis it was decided to create system which picks news, processes news texts and automatically groups them by content. The created system differs from other solutions in Lithuanian market that the users get dynamic news groups, which varies according to news. The created news aggregator's prototype works like this: according to predetermined parameters news are assembled periodically from different news sites, then news texts are processed. News are aggregated for the specified time period. System performs information processing and display functions parallel to each other.

Research was performed during the creation of the system prototype, which aimed to find the best way to process news texts, so that the information grouping would give meaningful categories. Experiment results led to conclusion that most meaningful news categories are obtained when texts consisting only of notional words are grouped. Moreover analysis of collected data for other uses and system future expansions was performed.

1. ĮŽANGA

1.1. Dokumento paskirtis

Šio dokumento paskirtis yra pateikti visą informaciją susijusią su kurtos sistemos - Lietuviško automatinio naujienų agregatoriaus prototipo realizacija, atliktais tyrimais ir jų rezultatais. Dokumentas sudarytas iš skyrių, kuriuose pateikiama atlikta mokslinių tyrimų, algoritmų ir įrankių analizė. Projektinėje dalyje pateikiama informacija apie sistemos realizacija: naudotus įrankius, realizuotus algoritmus. Eksperimentinėje dalyje aprašomas vykdytas tyrimas, sistemos plėtojimo perspektyvos ir kitos įžvalgos.

1.2. Darbo tikslai

Pagrindinis darbo tikslas yra sukurti prototipinę sistemą, kuri pasiūlytų patogų ir laiką taupantį būdą skaityti naujienas iš įvairių šaltinių. Sistemos tikslas - grupuoti Lietuviškas naujienas iš įvairių naujienų portalų, kad vartotojai efektyviau ir greičiau rastų pagal kategorijas suskirstytas konkrečias naujienas. Darbo akcentas yra automatinis naujienų grupavimas pasitelkiant mašininio mokymosi (angl. „Machine Learning“) algoritmus. Informacijos skirstymas ir grupavimas, o dar svarbiau ryšių tarp skirtingų informacijos vienetų išskyrimas ir pateikimas galutiniam vartotojui yra aktuali ir perspektyvi programinių produktų plėtojimo kryptis.

1.3. Santrauka

Kasdien sukuriamas didelis informacijos srautas, kurį be įvairių programinių įrankių būtų sunku aprėpti. Kaip vienas iš palengvinančių informacijos surinkimo ir pateikimo būdų yra „RSS“ (angl. „Rich Site Summary“) prenumeratos. Ši technologija skirta greitai ir saugiai surinkti bei pateikti naujausius naujienų įrašus iš įvairių naujienų šaltinių. Sistemos tikslas - grupuoti Lietuviškas naujienas iš įvairių naujienų portalų tam, kad vartotojai efektyviau ir greičiau rastų konkrečias tik juos dominančias naujienas. Gaunama informacija yra grupuojama į kategorijas, tačiau čia atsiranda daugybė vietos interpretacijoms: kokios turėtų būti grupės, kaip skirstyti ir susieti skirtingą informaciją.

Norint sukurti „protingą“ naujienų agregatorių pirma buvo atlikta grupavimo metodikų, algoritmų ir įrankių analizė. Pagal atliktą analizę nuspręsta kurti sistemą, kuri surenka naujienas, apdoroja naujienų tekstus bei automatiškai grupuoja naujienas pagal jų turinį. Sukurta sistema išsiskiria iš kitų Lietuvos rinkoje esančių sprendimų tuo, kad

vartotojui pateikiamos dinaminės naujienų grupės, kurios kinta priklausomai nuo pateikiamų naujienų. Sukurtas naujienų agregatoriaus prototipas pagristas tokius veikimu: pagal nustatytus parametrus periodiškai surenkamos naujienas iš įvairių naujienų portalų, surinktų naujienų tekstai yra apdorojami, o naujienos kaupiamos pagal pasirinkta laikotarpį. Sistemoje informacijos apdorojimo ir atvaizdavimo funkcionalumas vykdomas lygiagrečiai.

Sistemos prototipo kūrimo eigoje buvo atliktas tyrimas, kurio metu buvo siekiama išsiaiškinti kaip turi būti apdoroti naujienų tekstai, kad atlikus naujienų grupavimą būtų gautos prasmingos kategorijos. Tyrimo rezultatai parodė, kad prasmingiausios naujienų kategorijos gaunamos tuomet, kai grupuojami naujienų tekstai sudaryti tik iš prasminių žodžių. Taip pat buvo analizuojami kiti surinktų duomenų panaudojimo būdai ir tolesnės sistemos plėtojimo galimybės.

2. ANALITINĖ DALIS

Šiame skyriuje pateikta mokslinių publikacijų apžvalga naujienų agregatorių kūrimo ir tekstinių dokumentų grupavimo temomis. Taip pat pateikiama naujienų grupavimo algoritmų ir įrankiu apžvalga.

2.1. Kas yra naujienų agregatorius

Naujienų agregatorius yra internetinė paslauga arba paprasčiau - svetainė, kuri surenka informaciją iš daugybės šaltinių ir parodo ją viename lange. Naujienų agregatorius gali būti realizuotas daugybe formų: nuo paprastos kompiuterinės programos, naršyklės įskiepio, mobilios programėlės ar internetinio servizo.

Kadangi yra skirtingų naujienų agregatorių realizavimo formų taip pat logiška manyti, kad agregatoriai kuriami siekiant skirtingų tikslų. Straipsnyje „The Rise of the News Aggregator: Legal Implications and Best Practices“ autorė išskiria naujienų agregatorius į tokias grupes [1]:

- Įrašų agregatoriai (angl. „Feed Aggregators“).
- Specializuoti agregatoriai (angl. „Specialty Aggregators“).
- Vartotojų redaguojami agregatoriai (angl. „User-Curated Aggregators“).
- Tinklaraščio tipo agregatoriai (angl. „BlogAggregators“).

Įrašų agregatoriai yra skirti surinkti ir rodyti straipsnių antraštes, kelias pirmas straipsnių eilutes ir nuorodą į šaltinį. Specializuoti naujienų agregatoriai skirti rinkti naujienas, kurios atitinka konkrečią tematiką ar vietovę. Vartotojo redaguojami agregatoriai yra skirti rinkti naujienas iš šaltinių, kuriuos nurodo pats vartotojas. Tinklaraščio tipo agregatoriai skiriasi nuo kitų naujienų agregatorių informacijos pateikimo koncepcija: informacija surenkama apie konkrečią temą apdorojama ir pateikiama kaip įprastas internetinis tinklaraštis. Konkretesni agregatorių pavyzdžiai pateikiami sekančiame skyriuje.

2.2. Vyraujančios tendencijos rinkoje

Naujienų agregatoriai iškilo internetinėje erdvėje dėl šių priežasčių [2]:

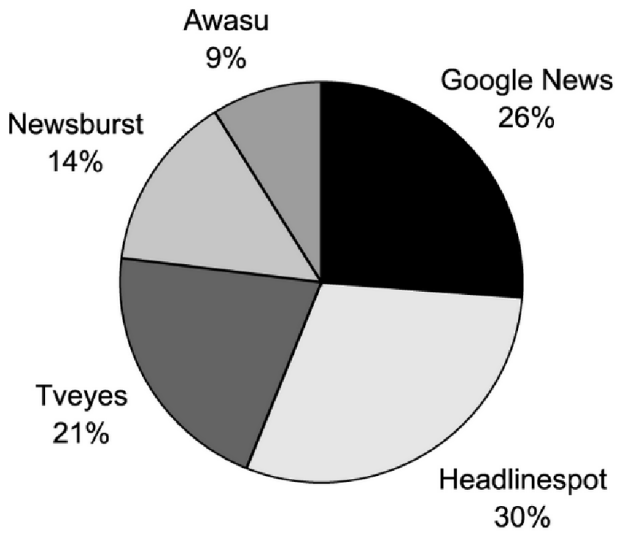
1. Vartotojai turi pilną gaunamo turinio kontrolę (prenumeratos gali būti lengvai ištrinamos, vartotojams nereikia pateikti elektrono pašto adreso ar kitos informacijos norint naudotis paslaugomis);
2. Sunku siųsti vartotojams brukalus (angl. „*spam*“) dėl to, kad vartotojai žino skaitomų naujienų šaltinių sąrašą.
3. Duomenys perduodami „RSS“ kanalais yra saugūs ir jais negalima perduoti virusų.
4. Duomenų perdavimas yra praktiškas, nes gaunamą informaciją galima panaudoti įvairiems tikslams: perduoti žinių kanalams, integruoti į dinamines bibliotekas.

Nors naujienų agregatoriai yra patogūs, tačiau vyrauja tendencija, kad mažiau išsivysčiusiose šalyse vartotojai dar neįsisavino laiką taupančių „RSS“ technologijų [2].

Tam, kad būtų atskleista, kokios tendencijos vyrauja, galima remtis 2006 metais atliktu tyrimu, kurio pagrindinis tikslas atskleisti naudotojų lūkesčius ir patirtį naudojant naujienų agregatorius [3]. Tyrime išskirti tokie naujienų agregatorių tipai:

1. Agregatoriai, kurie tik surenka informaciją ir ją pateikia neapdorotą („Headlinespot“, „GoogleNews“).
2. Agregatoriai, kurie surenka informaciją ir apdoroja atitinkamai pagal vartotojo reikalavimus („TVEyes“, „Newsburt“, „Awasu“).

Tyrimas atskleidė, kaip pasiskirsto vartotojai, pagal naudojamus naujienų agregatorius:



1 pav. Vartotojų pasiskirstymas pagal naudojamus agregatorius [3]

„Headlinespot“ (www.headlinespot.com) – skirtas JAV šaliai. Pateikia naujienas pagal temą, geografinę žymę. Galima skaityti kritikų atsiliepimus, pasirinkti naujienų šaltinius: internetinius blogus, laikraščius, televizijos transliacijas.

„GoogleNews“ (http://news.google.com) - pateikia naujienas iš viso pasaulio pagal kategorijas, kuriose naujienos suskirstytos pateikiant daugiausiai ryšių turinčias naujienas aukščiau. Įgalintas prisitaikymas prie vartotojų pageidavimų.

„Newsburst“ (www.newsburst.com) – pateikia priėjimą prie naujienų, internetinių tinklaraščių, orų pranešimų ir kitos informacijos. Vartotojai gali personalizuoti naujienų agregatorių sukurdami tik juos dominančias naujienų kategorijas.

„Awasu“ (www.awasu.com) – pateikiamas trimis versijomis: personalinė, profesionali, aukšto lygio. Personalinė versija yra nemokama. Sistema praneša apie naujienas iš pasirinktų vartotojo temų ir renka duomenis apie vartotojų perskaitytas naujienas.

„TVEyes“ (www.tveyes.com) – realiu laiku pateikia naujienas iš televizijų, radijo transliacijų. Vartotojai mėgstamus įrašus ar vaizdo klipus gali peržiūrėti sukurtuose archyvuose.

Taip pat svarbu atkreipti dėmesį į funkcijas, kurios svarbiausios vartotojui. Atliktame tyrime parodyta, kokiomis funkcijomis galima pritraukti interneto vartotojus (1 lentelė).

1 lentelė Vartotojui svarbiausios funkcijos naujienų agregatorių veikime

Naujienų agregatorių funkcijos	Procentinė naudotojų dalis, kuri funkciją nurodė kaip labiausiai pageidaujamą savybę
Išplėstinis paieškos funkcionalumas.	80 %
Patogi vartotojo sąsaja.	77,8 %
Kokybiški ir turintys gerą reputaciją šaltiniai.	75,6 %
Senų įrašų paieškos galimybė.	71,1 %
Patogus naršymo funkcionalumas.	68,9 %
Galimybė rodyti rezultatus chronologine tvarka.	64,4 %
Personalizavimo funkcionalumas.	62,2 %
Galimybė paslauga naudotis nemokamai.	57,8 %
Didelis kiekis išvedamų aktualių rezultatų.	53,3 %
Pranešimų paslauga.	48,9 %
Įrašų kita kalba vertimo funkcija.	46,7 %
Galimybė peržiūrėti informaciją naudojantis išmaniąją televizija.	40,0 %

Vartotojas naudodamasis naujienų agregatoriumi pirmiausia atkreipia dėmesį į paieškos funkcionalumą. Teikiama pirmenybė toms sistemos, kurios analizuoja ir susistemina naujienų srautus taip palengvinant naudotojams paiešką tarp daugybės naujienų įrašų.

Norint realizuoti sėkmingus projektus ar produktus svarbu žinoti, kokios rinkos tendencijos ir šakos iškils artimoje ateityje. Pavelo Marceukso atlikta analizė pateikė sąrašą IT verslo tendencijų, kurios bus populiarios 2020-aisiais [4]. Penktoje vietoje pateikta tendencija, kad iškils poreikis programinės įrangos, kuri galėtų apdoroti ir greitai analizuoti didelius kiekius informacijos. Susisteminta informaciją bus įtakinga ir vertinga priemonė prekinių ženklų formavime, tikslinėms auditorijoms ar įmonėms.

2.3. Situacijos Lietuvoje įvertinimas

Kiekvieno naudotojo įpročiai yra skirtingi, todėl kiekvienas informacijos skleidimo šaltinis turi savo naudotojų grupę. Informacinės visuomenes plėtros komiteto prie Susisiekimo ministerijos užsakymu 2011 metais buvo atliktas tyrimas, kurio rezultatai parodė augantį interneto naudotojų skaičių [5]. Tyrimo metu išsiaiškinta, kad internetu naudojasi 69 procentai respondentų. Taip pat buvo klausiama ką dažniausiai žmonės veikia internete. Net 47,8 procentai respondentų pažymėjo, jog internete skaito lietuviškus laikraščius bei žurnalus. Galima daryti išvadą, kad augant skaičiui žmonių, kurie naudojasi

internetu auga ir skaičius žmonių, kurie apie jiems aktualius įvykius ir naujienas informacijos ieško internete.

Lietuvos internetinėje erdvėje šiuo metu gyvuoja du pagrindiniai naujienų agregatoriai: „visosnaujienos.lt“ ir „www.glaustai.lt“. Pagrindinis šių agregatorių veikimo principas yra naujienų surinkimas iš įvairių šaltinių „RSS“ formatu bei pateikimas pagal tam tikrus kriterijus:

- suskirstymą į temas temas, kaip „Sportas“ „Mokslas ir IT“,
- įrašų naujumą,
- įrašų populiarumą.

Peržiūrėjus „visosnaujienos.lt“ sistemą matoma, kad svetainėse pateikiamos naujienų kategorijos yra statinės ar priklausomos nuo konkrečių naujienų šaltinių pateikiamos informacijos. Kaip privalumas – svetainėje pateikiamų kategorijų pavadinimų dydžiai vizualiai parodo kiek naujienų priskirta tai temai.

Svetainėje „www.glaustai.lt“ pateikiamos statinės naujienų kategorijos, tačiau galima temas išskleisti ir peržiūrėti vidines kategorijas. Didžiausiais minusas – kai kurios kategorijos neturi priskirtų naujienų ir tai sužinoma tik atidarius konkrečią kategoriją.

2.4. Problemos iškylančios kuriant naujienų agregatorių

Naujienos generuojamos kiekvieną dieną, todėl gaunami dideli informacijos kiekiai. Iškyla problema – kaip atsirinkti tik svarbiausią informaciją. Realizuotose sistemose naujienų įrašai klasifikuojami pagal jų sukūrimo ar publikavimo datą, o įrašų filtravimas galimas tik pagal naudotojo nurodytus raktažodžius arba taisykles. Dėl šių trūkumų naudotojai gauna didelius kiekius informacijos, kurie dažniausiai neturi bendrų ryšių. Kuriami naujienų agregatoriai siekiantys kuo efektyviau susisteminti informaciją prieš jai pasiekiant galutinį naudotoją.

Autorių teisių pažeidimo problema taip pat iškyla kuriant naujienų agregatorių. Pagal Autorių teisių ir gretutinių teisių įstatymo ketvirtą skirsnį 24 straipsnį leidžiama viešai skelbti ar padaryti viešai prieinamus išleistus straipsnius aktualiomis ekonomikos, politikos ar religijos temomis, taip pat analogiško pobūdžio transliuojamus kūrinius, jeigu autoriai ar kiti tų kūrinių autorių teisių subjektai nėra uždraudę taip naudoti kūrinius ir jeigu nurodomas šaltinis, įskaitant autoriaus vardą [6]. Remiantis šiuo įstatymo straipsniu, kuriant Lietuvišką naujienų agregatoriaus prototipą nebus pažeidžiamos autorių teisės.

Taip pat legalumo principas nagrinėjamas straipsnyje „*The Rise of the News Aggregator: Legal Implications and Best Practices*“ [1], kuriame pateikiama geriausia praktika kuriant naujienų agregatorių nepažeidžiant autorių teisių:

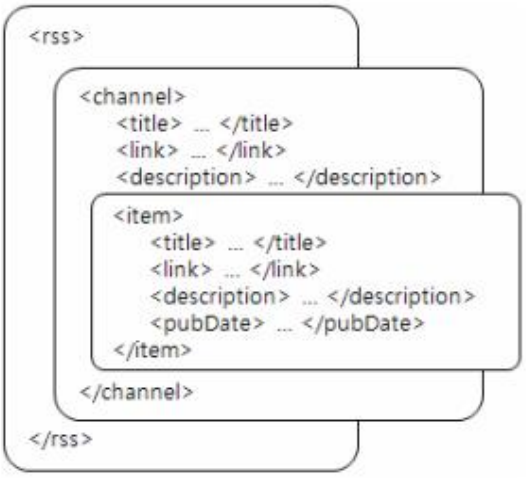
- Naudoti tik tas naujienų antraštes, kurios yra reikalingos identifikuoti turiniui ar atlikti veiksmus, ir nekopijuoti naujienų turinio.
- Naudoti tik tuos straipsnius, kurie yra susiję.
- Matomoje vietoje pateikti straipsnio šaltinio pavadinimą.
- Pateikti nuorodas į šaltinį.

2.5. Techniniai aspektai

Šiame skyriuje pateikiama informacija apie techninius naujienų agregatoriaus kūrimo aspektus. Remiantis kitų mokslinių darbų medžiaga surinkta ir pateikiama apžvalga apie duomenų, reikalingų naujienų agregatoriaus prototipo kūrimui, surinkimą bei apdorojimą.

2.5.1. Technologijos

Naujienų agregatoriaus veikimas įgyvendintas „RSS“ („Rich Site Summary“)technologija. RSS yra informacijos, pateikiamos „XML“ formatu, surinkimo iš internetinių portalų technologija. Naujienų įrašai pagal pareikalavimą yra parsisiunčiami iš žinomų šaltinių. Kaip pavaizduota 2 pav., dažnai parsisiunčiama informacija susideda iš tokių dalių: įrašo pavadinimo, nuorodos, aprašymo, publikavimo datos [7].



2 pav. Pavyzdinė RSS įrašo struktūra [7]

Gaunamas srautas naujienų įrašų dažnai turi bendrą temą, kurią nurodo naujienų autoriai. Temos taip pat bendros kategorijos, kurioms priskiriami naujienų įrašai, dažnai yra abstrakčios ir toks naujienų pateikimo būdas neleidžia greitai peržiūrėti labiausiai

dominančias naujienas. Siekiant susisteminti gaunamus naujienų srautus kuriami „protingi“ algoritmai, kurių pagrindinis darbas susisteminti gaunamus naujienų įrašus. Detaliau apie naudojamus algoritmus ir grupavimo įrankius nagrinėjama skyriuje 6 „Algoritmų ir įrankių analizė“.

2.5.2. Duomenų paruošimas

Kuriant autonomiškas sistemas vienas iš pasiruošimo darbų yra tinkamai pateikti duomenų rinkinius. Naujienų straipsniai „XML“ formatu parsisiunčiami iš įvairiausių šaltinių yra suprantami tik žmonėms. Tam, kad mašina galėtų sukurti ryšius tarp duomenų rinkinių, klasifikuoti ar sisteminti informaciją, pradiniai duomenys turi būti apdoroti. Pateikiami pagrindiniai pradinių duomenų apdorojimo metodai [8] :

1. Naudoti „žodžių maišo“ (angl. „bag-of-words“) modelį. Šiame modelyje nepaisoma skyrybos ženklų ar žodžių eiliškumo, o tekstai yra išskaidomi į atskirus žodžius ir saugomas tų žodžių pasikartojimo dažnis konkrečiame dokumente.
2. Pašalinti dažnai pasikartojantys žodžiai, ko pasėkoje tekste paliekami tik prasminiai žodžiai.
3. Vykdomas kamienizavimas - sukuriamas žodynas ar komponentas susiejantis panašių reikšmių žodžius (pvz. tą patį žodžio kamieną turintys žodžiai reprezentuoja bendrą prasmę).

2.6. Algoritmų analizė

Svarbiausios naujienų agregatoriaus dalys būtų:

- komponentas, atsakingas už informacijos surinkimą,
- komponentas, kuris atsakingas už įrašų klasifikavimą ir klasterizavimą.

Tolesniuose skyriuose apžvelgiami algoritmai ir įrankiai, kurie yra arba buvo naudojami naujienų agregatorių kūrimo ar tobulinimo procesuose. Kadangi mašininio mokymosi algoritmų realizacijų yra daug, todėl šiame darbe apžvelgiami algoritmai, kurie yra daugiausiai paplitę arba, kurių realizacijos idėjos yra naujovė informacijos grupavimo srityje ir įvairių programinių įrankių kūrime.

2.6.1. „K-means“ algoritmas

Kaip vienas iš plačiausiai žinomų yra „K-means“ algoritmas. Šis algoritmas yra naudojamas duomenų grupavimui. Algoritmo standartas pirmą kartą pasiūlytas 1965 metais, o pirmą kartą moksliniame leidinyje paminėtas 1966 metais [9]. Šio algoritmo veikimas priklauso nuo duomenų rinkinio dydžio. Priklausomai nuo realizacijos, šis

algoritmas reikalauja vartotojo įsikišimo, kurio metu būtų nurodytas klasterių kiekis. Algoritmo veikimas yra priklausomas nuo nustatyto centrų kiekio *k*. Taip pat jei nustatytas didelis duomenų rinkinys, algoritmo veikimas gali užimti labai daug laiko [10]. Pagrindinis „K-means“ algoritmas ieškant *k* klasterių [11]:

1. Pasirenkamas *k* kiekis taškų, kurie bus pradiniai centroidai, Visi taškai priskiriami artimiausiems centroidams.
2. Perskaičiuojami centroidai kiekvienam klasteriui.
3. Kartojamas 2 ir 3 žingsniai iki tol, kai centroidai nebekinta.

Kadangi šis klasifikavimo algoritmas yra vienas iš populiariausių, todėl yra realizuota ne viena šio algoritmo variacija. Bradley‘us ir Fayyad‘as pateikė patobulintą „K-means“ algoritmą [12]. Jų algoritmo esmė – atsitiktine tvarka skaidyti duomenis į 10 atskirų poaibių. Vėliau, kiekvienam poaibiui atliekamas klasifikavimas panaudojant įprastą „K-means“ algoritmą. Surasti centroidai kiekviename poaibyje vėl perduodami algoritmui, kurio rezultatas parodo galutinę centroidų vietą.

2.6.2. Hierarchinis įrašų grupavimas

Hierarchinis įrašų grupavimas skiriasi nuo kitų algoritmų savo gaunamo rezultato struktūra: gražinamas genealoginis medis, kuris turi vieną bendrą visus klasterius apimančią viršūnę. Išskleidžiant viršūnes nuo pradinės išplečiami pavieniai unikalūs klasteriai.

Yra du pagrindiniai hierarchinio grupavimo metodai [11]:

1. Kiekvieną dokumentą laikyti individualiu klasteriu ir kiekviename žingsnyje sujungti daugiausiai panašumų turinčius klasterius ar klasteriu grupes. Būtinas grupių panašumo apibrėžimas.
2. Visus dokumentus laikyti viename klasteryje ir kiekviename žingsnyje skaidyti klasterį į grupes kol jose lieka tik po vieną unikalų įrašą. Skaidant būtina apibrėžti sąlygas kada skaidyti klasterį ir kaip atlikti grupių išskyrimą.

Pirmasis metodas yra taikytinas dažniau. Metodo realizavimo idėja apima tokius veiksmus:

1. Skaičiuojama panašumas tarp visų klasterių. (skaičiavimams galima naudoti panašumo reikšmių matricą, kurios ij^x reikšmė žymi panašumo reikšmę tarp i^x ir j^x klasterių).
2. Suliejami panašiausi (arčiausiai esantys) klasteriai.
3. Atnaujinama panašumo reikšmių matrica.

4. Kartojami 2 ir 3 žingsniai iki kol lieka tik pavieniai klasteriai.

Šio algoritmo privalumas yra tai, kad informacija pateikiama struktūriškai sugrupuota. Atlikus grupavimą gražinamas hierarchinis medis, kurio viršūnės atitinka informacijos grupes.

2.6.3. Automatiškas dokumentų grupavimas naudojant žodžių rinkinius

Standartiniai ir gerai žinomi algoritmai kaip „K-means“ gali būti naudojami atliekant dokumentų klasterizavimą, tačiau jie neatitinka specifinių reikalavimų: efektyvaus darbo su didelės apimties duomenimis, lengvo naršymo ir prasmingų grupių išrinkimo.

Dokumentų klasifikavimą galima atlikti naudojant žodžių rinkinius. Žodžių rinkiniai gali butį dažnai tekste pasikartojantys žodžiai, kurie išreiškia bendrą idėją. Žodžių rinkinio pavyzdys: [tigras, zoologijos, sodas] kuris gali apibūdinti naujienas, kuriose kalbama apie tigrus laikomus zoologijos soduose. A. Sharma ir R. Dhir pateikė savo požiūrį į dokumentų grupavimą: vietoje tiesioginio ar remiantis pasikartojančiais terminais dokumento lyginimo galima dokumentus grupuoti remiantis uždarais žodžių rinkiniais (angl. „*closed word sets*“) [13]. Pateiktą požiūrį autoriai pagrindžia tokia algoritmo veikimo idėja - pirmiausia kiekviename dokumente ieškoma globalių dažnai pasikartojančių žodžių rinkinių, toliau surastiems žodžių rinkiniams suformuojamas pradinis klasteris, kuriam priskiriami dokumentai, savyje turintys atitinkamus žodžių junginius. Po to klasteriai, kurie turi sutampančius dokumentų rinkinius, yra išskaidomi pagal reitingavimo funkciją. Atliktų veiksmų rezultatas gražina unikalius klasterius, kuriuose saugomi panašius žodžių junginius turintys dokumentai.

Algoritmo veikime naudojami tik prasmingi žodžiai, taip suspaudžiamas pradinis tekstas ir sumažinama skaičiavimų apimtis bei laikas.

2.6.4. Istorijos sekimo – panašių naujienų susiejimo pagal laiko žymę algoritmas

Kuriant naujienų agregatorius vienas iš svarbiausių aspektų pateikiant informaciją yra susijusių naujienų pateikimas pagal pasirodymo laiką. Straipsnyje „*Story tracking: linking similar news over time and a cross languages*“ autoriai pateikia tokias naujienų susiejimo pagal laiką taisykles [14]:

- Ta pati istorija gali būti sudaryta iš kelių informacijos klasteriu.
- Jei naujas klasteris yra panašus į bent vieną iš istorijai priskirtų klasteriu, tai tikėtina, kad naujas klasteris yra istorijos tęsinys.

- Unikalios ar per septynias dienas nesusietos istorijos yra priskiriamos naujai sukuriamam klasteriui.

Realizuojant algoritmo veikimą svarbiausia dalis nustatyti ar naujas klasteris yra susijęs su istorija, tai galima atlikti skaičiuojant panašumo laipsnį. Panašumo laipsnis gaunamas atliekant lyginimus tarp naujai sukurto klasterio su visais konkrečios istorijos klasteriais pradedant nuo vėliausiai pridėto, vėliau dauginant iš „mažinimo“ koeficiento. Mažinimo koeficientas padeda išlaikyti panašumo faktoriaus reikšmę intervale 0 iki 1.

Teorinės panašumo ribos yra tokios: 0 – istorijos yra nesusijusios, 1 – istorijos yra labai susijusios. Jei pagal lygtį apskaičiuotas rezultatas peržengia 0,5 slenkstį tai laikoma, kad nagrinėjami naujienų klasteriai yra susiję.

2.6.5. RSS įrašų klasifikavimas naudojantis Dirbtine imunine sistema

Dirbtinė imuninė sistema (angl. „Artificial Immune System“) naudojama autonominiai navigacijai, kompiuterinei apsaugai. Dirbtinė imuninė sistema (AIS) yra viena iš mašininio mokymosi sistemų, kuri suteikia galimybę greitai prisitaikyti ir evoliucionuoti priklausomai nuo atliekamų užduočių. Naujienų agregatorių kūrime tokios sistemos naudojamos siekiant nustatyti, kurie naujienų įrašai iš turimų rinkinių yra susiję tarpusavyje. AIS sistemos idėja buvo iškelta dar 1980 metų viduryje, bet plačiai paplito tik devintame dešimtmetyje [15].

AIS sistemos veikimo metodai išdėstyti straipsnyje „*Classifying RSS Feeds with an Artificial Immune System*“: atėjus naujam naujienos įrašui sistema generuoja elementą, kuris reprezentuos naujienos straipsnį [15]. Sugeneruoti elementai savyje saugo atsitiktine tvarka parinktus žodžius iš naujienos antraštės ar aprašymo. Kadangi negalima automatiškai atrinkti žodžių, kurie geriausiai atvaizduoja naujienos turinį, sistema kiekvienam straipsniui generuoja kelis skirtingus elementus. Geresniam algoritmų veikimui naudojama morfologija bei nereikšmingų žodžių išmetimas. Turint sugeneruotus elementus kiekvienam straipsniui galima kurti palyginimo algoritmus. AIS sistemoje elementų palyginimas priklauso nuo šių faktorių:

1. Procentinis skaičius pasikartojančių žodžių tarp skirtingų straipsnių.
2. Procentinis skaičius pasikartojančių elementų tarp skirtingų straipsnių.
3. Straipsnio pasikartojimo skaičius.

Jei lyginamų straipsnių faktorių parametrai peržengė nustatytą slenkstį, tai šie straipsniai laikomi susijusiais.

2.6.6. Algoritmų palyginimas

Atlikus algoritmų analizę galima išskirti pagrindinius algoritmus, kurie gali būti naudojami naujienų grupavimui: „K-means“ ir „Hierarchinio įrašų grupavimo. Algoritmų palyginimas[16]:

- Didėjant duomenų kiekiui „Hierarchinio įrašų grupavimo“ efektyvumas mažėja ir veikimo laikas didėja.
- Didėjant duomenų kiekiui tikslumui išsaugoti tinkamesnis yra grupavimas pagal žodžių rinkinius.
- „K-means“ algoritmo veikimo laikas taip pat didėja didėjant duomenų imčiai, tačiau lyginant su „Hierarchinio įrašų grupavimo“ algoritmu, efektyvumas išlieka aukštesnis.
- „Hierarchinis įrašų grupavimo“ algoritmas pateikė tikslesnius rezultatus lyginant su „K-means“ algoritmu.

Kiti nagrinėti algoritmai gali būti naudojami naujienų istorijų sekimui laike ar filtravimui atlikti.

2.7. Įrankių analizė

Apžvelgiami įrankiai, skirti klasifikuoti ir klasterizuoti tekstus, kuriuose naudojami ir realizuoti aukščiau paminėti algoritmai. Įrankiai išrinkti pagal populiarumą ar naudojimą kituose moksliniuose tyrimuose. Taip pat aprašomas įrankis skirtas žodžių kamienizavimui.

2.7.1. Įrankis „carrot^2“

„Carrot^2“ yra atviro kodo paieškos rezultatų klasterizavimo variklis [17]. Pagrindinė paskirtis – automatiškai organizuoti nedidelius rinkinius duomenų, paieškos rezultatų ir kita informaciją į temines kategorijas. Šis įrankis 2004 metais buvo apdovanotas specialiu prizu „EASA“ konkurse, taip pat plačiai naudojamas mokslinėse publikacijose [18]. Klasterizavimui naudoja šiuos algoritmus:

- „Lingo“ - šis algoritmas naudoja dažniausių frazių paiešką ir semantinio indeksavimo būdą atskiriant duomenis į grupes. Klasterių pavadinimams sudaryti ieško prasminių frazių tekste [19]. Pirma bandoma sudaryti prasmingą klasterio pavadinimą, pagal kurį bus atrenkami tekstai.
- „STC“ - algoritmas analizuoja pasikartojančius žodžių rinkinius tekstuose, pagal juos skaičiuojamas dokumentų panašumas. Pasikartojantys rinkiniai randami

skaidant dokumentų tekstus į blokus, pasikartojantys blokai saugomi medžio tipo struktūroje [20].

- „Bisecting K-means“ - algoritmas klasterizuoja tekstynus į grupes, kuriose vienas dokumentas gali priklausyti tik vienai grupei. Grupių pavadinimai sudaryti iš pavienių žodžių, kurie buvo daugiausiai pasikartoje tekstynuose.
- “ „Lingo3G[1]“ - algoritmo veikimas panašus kaip ir „Lingo“, tačiau įgalintas hierarchinis klasifikavimas.

2.7.2. Įrankis „LingPipe“

„LingPipe“, sukurtas2003 m. kompanijos „Alias-i“. Šis įrankis skirtas analizuoti tekstą naudojantis kompiuterine lingvistika. Šis įrankis projektuotas atlikti tokias užduotis [21]: ieškoti žmonių, pavadinimu ir vietų naujienų srautuose, automatiškai klasifikuoti „Twitter“ paieškos rezultatus, siūlyti rašybos klaidų taisymus.

Įrankis buvo projektuotas norint pasiekti kuo didesnio efektyvumo, įgyvendinti daugkartinį panaudojamumą ir išplėčiamumą. Kūrėjų svarbiausi įrankio bruožai:

- JAVA programavimo kalbos palaikymas.
- Daugiakalbiškų tekstų palaikymas.
- Mokymosi atlikti naujas užduotis su naujais duomenimis.
- Mokymas internetu.
- Saugus gijų, sinchronizacijos, koduočių palaikymas.

2.7.3. Įrankis „Apache Mahout“

Atviro kodo „Apache Mahout“ įrankis yra skirtas efektyviai apdoroti didelius kiekius informacijos. Siūlomos tokios pagrindinės duomenų analizės funkcijos: filtravimas, klasifikavimas, grupavimas [22]. Klasifikavimui ir grupavimui naudojami tokie algoritmai kaip atsitiktiniai miškai (angl. „Random forest“) – kolektyvinis mokymosi metodu paremtas klasifikavimo algoritmas. Klasifikavimui sudaromi medžiai, kiekvienas medis apsprendžia kuriam klasteriui priskiriami duomenys [23] „K-means“ klasifikavimo algoritmas ir kiti.

[1] Lingo3G prieinamas tik komercinėje versijoje.

2.7.4. Įrankis „Weka“

„Weka“ yra atviro kodo įrankis skirtas mašininio mokymosi ir duomenų gavybos užduotims atlikti [24]. Įrankyje naudojami algoritmai gali būti pritaikomi tiesiai duomenų rinkiniams arba naudojamai JAVA programavimo kalba parašytuose programose. „Weka“ įrankį sudaro tokios funkcijos: duomenų išankstinis apdorojimas, klasifikavimas, grupavimas, asociacijų taisyklių sudarymas, ir vizualizacija.

Duomenų klasifikavimui naudojami tokie algoritmai [25]:

- „Cobweb“ yra didėjančios hierarchinės sistemos konceptualus grupavimas. Algoritmo autorius prof. Douglas H. Fisher.
- DBSCAN algoritmas, iškilęs 1996 metais. Turint duomenų kiekį pažymėta plokštumoje algoritmas grupuoja taškus, kurie yra arčiausiai vienas kito.
- „Pirmas toliausiais“ algoritmas (angl. FARTHEST FIRST ALGORITHM) yra „K-means“ algoritmo variacija, kurioje kiekvieno klasterio centrą patraukia toliau nuo kitų klasterių centrų.
- „OPTICS“ algoritmas, kuris ieško klasterių pagal duomenų tankį erdvėje.
- „K-means“ algoritmas, kuris klasifikuoja objektus į apibrėžtus duomenų pasiskirstymo centrus.

2.7.5. Įrankis „Snowball“

„Snowball“ yra maža tekstinių duomenų apdorojimo kalba kurta kamienizavimo algoritmų realizacijai ir informacijos paieškai. Kamienizavimui palaikomos kalbos: Anglų, Prancūzų, Ispanų, Portugalų, Italų, Rumunų, Vokiečių, Švedų, Norvegų, Danų, Rusų, Suomių. Taip pat galima plėsti įranki pridedant kitų kalbų kamienizavimo algoritmus. [26].

Vienas iš šio įrankio panaudojimo pavyzdžių Lietuviškose projektuose yra „KąVeikiaValdžia.lt“. Realizuojant šią sistemą dokumentų paieškai naudojamas serveris – „ApacheSolr“, kuris leidžia greitai atlikti paieška tarp didelių dokumentų apimčių. Taip pat „Snowball“ įrankis praplėstas taip, kad suprastu Lietuviškų žodžių formas dėka kodo sukurto Žygimanto Medelio, M. Petkevičiaus, Tomo Krilavičiaus [27].

3. SIŪLOMAS SPRENDIMAS

Pagal atliktą analizę siūlomi tokie sistemos realizacijos sprendimai:

- Pradinių duomenų surinkimui naudojama RSS technologija.
- Tikslesnei duomenų analizei surinkti pilnus naujienos tekstus, pasitelkiant informacija gautą iš RSS srauto.
- Naujienų kategorijos turi būti sudaromos automatiškai pagal naujienų tekstinę informaciją.

Naujienų kategorijų generavimui ir naujienų grupavimui pasitelkiamas „Carrot^{2}“ įrankis. Šis įrankis pasirinktas nes jis yra plačiai naudojamas kituose mokslinio pobūdžio tyrimuose bei šiuo įrankiu lengva naudotis ir pateikiama informatyvi naudojimosi instrukcija. Tam, kad naujienų grupavimas vyktu greičiau ir tiksliau siūlomi tokie naujienų tekstų mažinimo metodai:

- Tekstai apdorojami pašalinant bendruosius žodžius, paliekant tik prasminius žodžius.
- Atliekamas žodžių kamienizavimas taip tekste sujungiami panašią prasmė turintys žodžiai.
- Atliekamas prasminių žodžių rinkinių išskyrimas.

4. PROJEKTINĖ DALIS

4.1. Sistemos pagrindinis funkcionalumas ir veikimo principas

Pagrindinės sistemos funkcijos yra surinkti antraštes iš Lietuviškų naujienų portalų, automatiškai grupuoti naujienas pagal tematiką ir pateikti sukurtame tinklapyje su nuorodomis į originalius straipsnius.

Sukurtas naujienų agregatoriaus prototipas pagrįstas tokiu veikimu: pagal nustatytus parametrus periodiškai surenkamos naujienas iš įvairių naujienų portalų, surinktų naujienų tekstai yra apdorojami, o naujienos kaupiamos pagal pasirinkta laikotarpį. Sistemoje informacijos apdorojimo ir atvaizdavimo funkcionalumas vykdomas lygiagrečiai.

4.2. Reikalavimų analizė

Sistemos kūrimo pradžioje buvo atlikta reikalavimų analizė, kurioje buvo išskirtos pagrindinės sistemos savybės:

- Gebėti išskirti naujienų kategorijas.
- Grupuoti naujienų tekstus.

Įprastai vartotojai, naujienas skaitantys internete, domisi tik maža dalimi pateikiamų įrašų, todėl naujienų agregatorius turi turėti realizuotas galimybes skaidyti ir rūšiuoti informaciją į atskirus blokus. Apibendrinus ir atsižvelgiant į antrame skyriuje (2 Analitinė dalis) atliktą analizę, sukurta sistema aprėpia pagrindinius bruožus:

- Saugoma visa informacija apie naujienos šaltinį (2.4 Problemos iškylančios kuriant naujienų agregatorių);
- Naujienos tekstų duomenys apdorojami paliekant tik svarbiausią informaciją (2.5.2 Duomenų paruošimas);
- Naujienų grupavimui pasitelkiami plačiai žinomi algoritmai ar įrankiai (2.6 Algoritmų analizė, 2.7 Įrankių analizė).

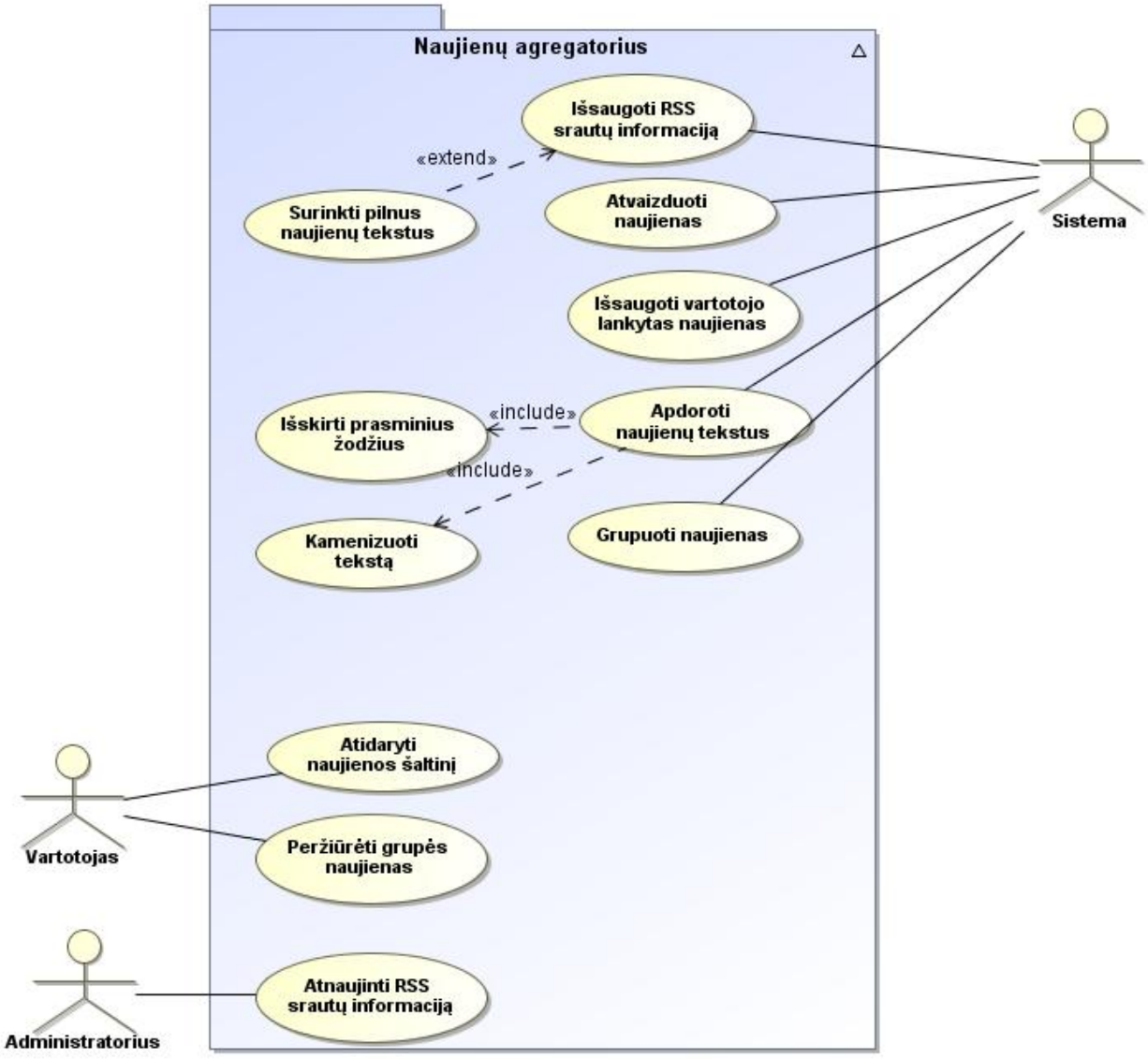
4.2.1. Nefunkciniai reikalavimai

Sistemos nefunkciniai reikalavimai :

- Naujienų kategorijos turi būti pateiktos patogiai.
- Kategorijos turi parodyti kiek susijusių naujienų jos turi.
- Netrukdomas naršymas tarp pateiktų naujienų.
- Pateikiamos naujienos išrikiuojamos pagal naujumą.
- Galima peržiūrėti pasirinktos kategorijos vidines kategorijas.

4.3. Panaudos atvejų diagrama

Atsižvelgus į išskirtus reikalavimus buvo parengta panaudos atvejų diagrama (2 pav.). Iškirti trys pagrindiniai aktoriai: vartotojas, administratorius, sistema.



3 pav. Panaudos atvejų diagrama

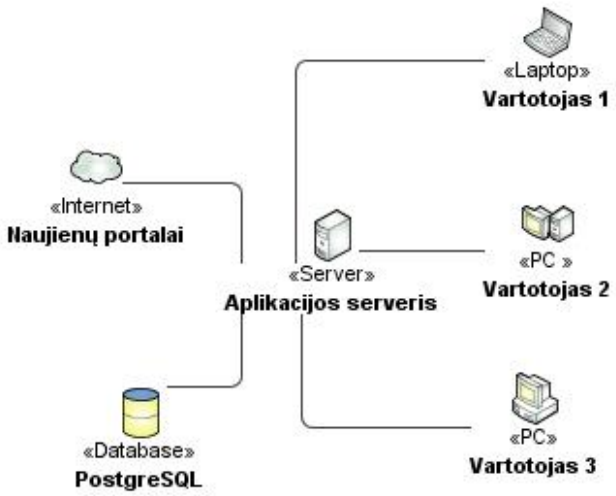
Daugiausiai panaudos atvejų priskirta sistemai, kadangi visas informacijos surinkimo ir apdorojimo periodinis procesas turi vykti automatiškai, be žmogaus įsikišimo. Sistemos vartotojas gali peržiūrėti sistemos veikimo rezultatus – sugrupuotas naujienas. Sistemos administratorius gali redaguoti naujienų šaltinių informaciją (pridėti naujų šaltinių, atnaujinti jau esamų šaltinių informacija).

4.4. Sistemos prototipo projektavimas

Sukurtos sistemos specifikacija pateikiama šiomis diagramomis: sistemos išdėstymo, sistemos statinio ir duomenų vaizdo. Tai pat aprašomi pagrindiniai informacijos apdorojimo algoritmai.

4.4.1. Išdėstymo vaizdas

Reikalinga techninė įranga, kurioje sistema bus išdėstyta ir veiks, pateikiama žemiau esančioje diagramoje.



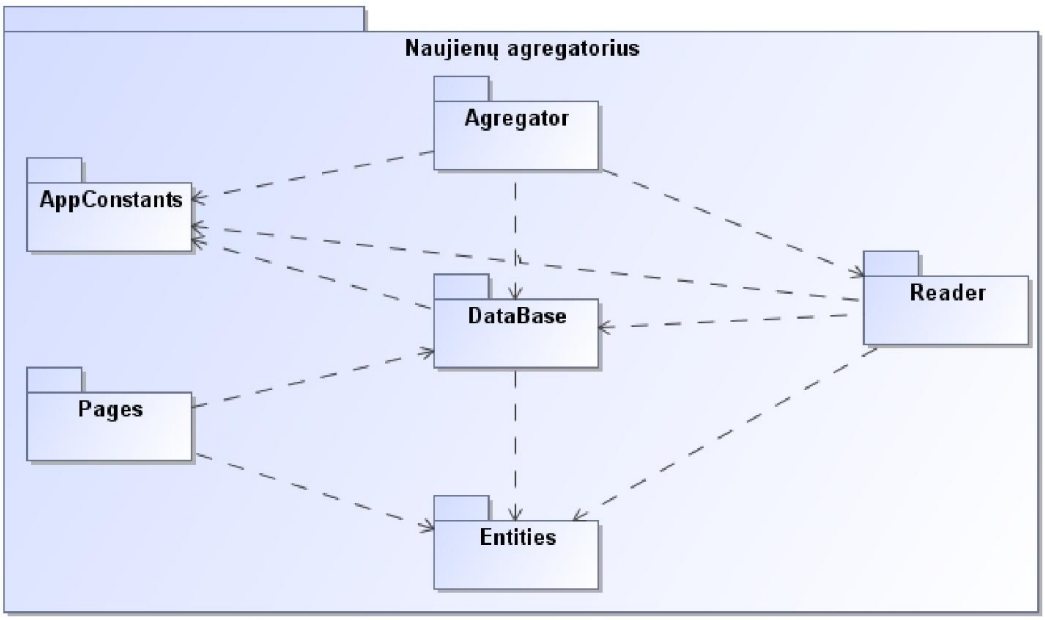
4 pav. Sistemos išdėstymo vaizdas

Svarbūs faktai apie sukurtą sistemą:

- Sukurtas sistemos prototipas gali vėliau išaugti į galutinę išplėstą sistemą.
- Nenaudota jokių statinių duomenų rinkinių įtakojančių gaunamus sistemos rezultatus.
- Dalis sistemos funkcionalumo realizuojama pasinaudojant jau sukurtais komponentais.
- Pagrindinis apribojimas norint, kad sistema pateiktų naujausią informaciją yra pastovi ir užtikrinta interneto prieiga.

4.4.2. Sistemos statinis vaizdas

Sistema suskaidyta į septynis paketus aukščiausiame lygyje, kuriuose realizuotas skirtingas funkcionalumas, bei visi paketai glaudžiai susiję vienas su kitų (5 pav.).



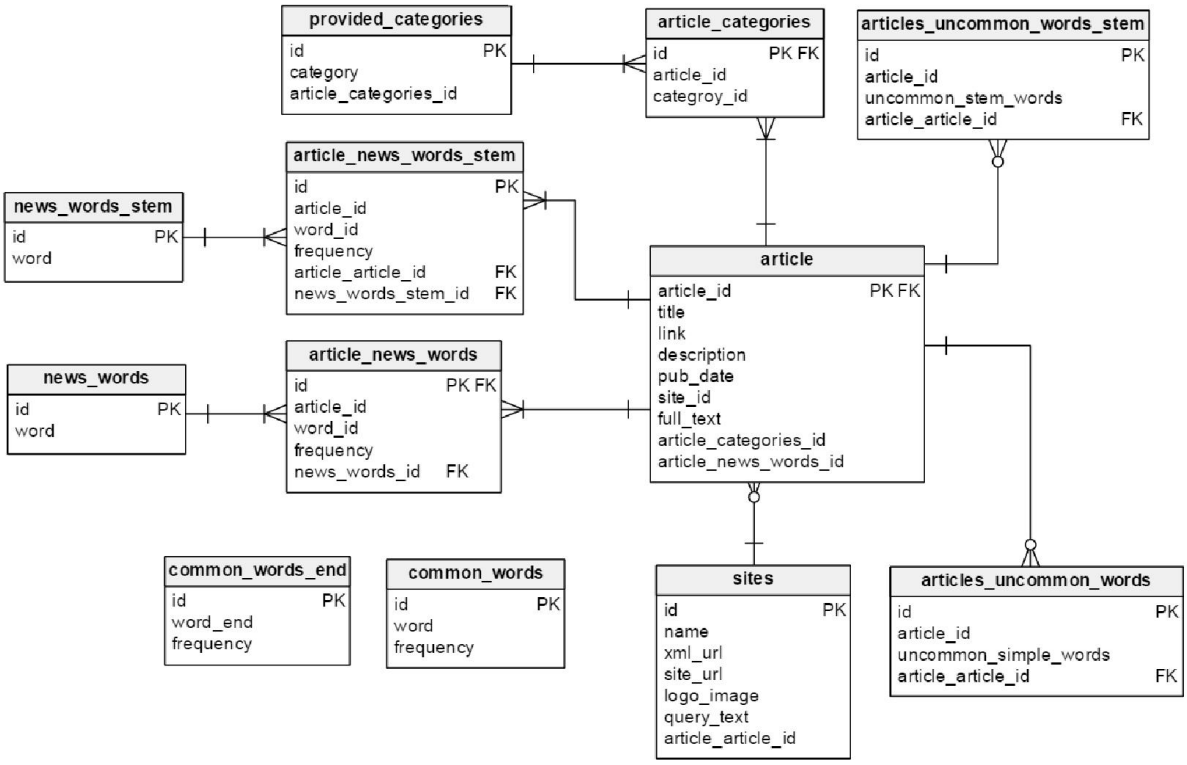
5 pav. Sistemos statinis vaizdas

Sistemą suskaldyta į šiuos paketus:

- „AppConstants“ pakete saugomos visos sistemoje naudojamos konstantos.
- „DataBase“ pakete realizuotos visos funkcijos reikalingos darbui ir bendravimui su duomenų baze.
- „Entities“ pakete laikomi sukurti ir naudojami duomenų tipai.
- „Reader“ pakete realizuotos funkcijos atsakingos už duomenų skaitymą, apdorojimą ir validavimą iš „RSS“ pateikiamų informacijos srautų.
- „Pages“ pakete saugomos klasės reikalingos internetinio naujienų agregatoriaus puslapių veiklai užtikrinti ir duomenų atvaizdavimui.
- „Agregator“ paketas atsakingas už naujienų pilnų tekstų apdorojimą ir naujienų grupavimą.

4.4.3. Duomenų vaizdas

Žemiau pateikiamas duomenų bazės modelis, kuris sukurtas sistemoje naudojamai informacijai saugoti.



6 pav. Duomenų bazės schema

Duomenų bazės lentelių aprašymas pateikiamas žemiau esančioje lentelėje.

2 lentelė Duomenų bazės lentelių aprašymas

Duomenų bazės lentelė	Aprašas
article	Skirta saugoti informaciją apie naujieną.
article_categories	Skirta saugoti informaciją apie naujienos kategorijas pateiktas kartu su naujienos informacija.
provided_categories	Skirta saugoti informaciją apie gautas kategorijas iš naujienų šaltinių.
site	Skirta saugoti informaciją apie naujienų šaltinius.
article_uncommon_words	Skirta saugoti informaciją apie naujienos tekstą gautą pašalinus dažniausius žodžius.
article_uncommon_words_stem	Skirta saugoti informaciją naujienos kamienizuotą tekstą su pašalintais dažniausiais žodžiais.
common_words	Skirta saugoti viso naujienų srauto žodžių žodyną.
common_words_end	Skirta saugoti viso naujienų srauto žodžių galūnių derinius ir jų pasikartojimą.
article_news_words	Skirta saugoti prasminių naujienos žodžius ir jų dažnį naujienoje.
news_words	Skirta saugoti prasminius žodžius visame naujienų sraute.
article_news_words_stem	Skirta saugoti prasminių kamienizuotų žodžių formas ir jų dažnį naujienoje.
news_words_stem	Skirta saugoti prasminių kamienizuotų žodžių formas visame naujienų sraute.

4.4.4. Bendras sistemos veikimo aprašymas

Sistemoje saugomos informacijos gavimas ir atnaujinimas vyksta periodiškai, kurio dažnumas apsprendžiamas pagal parametrą (naujienų atnaujinimo periodiškumas yra skaičiuojamas valandomis, pradinė reikšmė 1 valanda). Informacijos atvaizdavimas ir apdorojimas vyksta lygiagrečiai. Sistema inicijuoja informacijos atnaujinimo procedūrą, kuri susideda iš šių fazių:

1. Inicijuojamas naujienų atnaujinimas. Iš sistemoje saugomų naujienų portalų informacijos gaunamas RSS srautas, kuriame nuskaitoma informacija apie naujienas.
2. Inicijuojama naujienų dažniausių žodžių analizė ir dažniausių galūnių analizė.

3. Inicijuojamas naujienų tekstų mažinimas, atliekant nereikšminių žodžių šalinimo, prasminių žodžių analizę ir kamienizavimą.

Vartotojas mato visą atnaujintą informaciją, kuri yra saugoma duomenų bazėje. Kadangi informacijos atnaujinimas ir atvaizdavimas gali vykti lygiagrečiai - funkcionalumas, atsakingas už informacijos atvaizdavimą, yra netrikdomas.

4.4.5. Naujienų surinkimas ir saugojimas

Atsižvelgiant į skyriuje **Error! Reference source not found.** „**Error! Reference source not found.**“ pateiktą informaciją buvo realizuotas pradinių duomenų surinkimas. Pirmiausia rankiniu būdu buvo išrinkti pradiniai informacijos šaltiniai – Lietuviškų naujienų portalų RSS nuorodos. Pradinis šaltinių sąrašas susideda iš 33 RSS srautų, kuriuos sudaro 12 Lietuviškų naujienų portalų (0

Priedai. 3 lentelė).RSS srautų yra daugiau, nes kai kurių portalų RSS srautai skirstomi į sritis, kaip „Mokslas“ ar „Verslas“. RSS sraute informacija pateikiama XML formatu. Apie naujieną galima išskirti tokią informacija:

- naujienos pavadinimas,
- nuorodą į originalų šaltinį,
- naujienos trumpas aprašymas,
- nuoroda į komentarų puslapį,
- publikavimo data,
- naujienos kategorijos (priklauso nuo šaltinio).

Gaunamas tik dalinis naujienos tekstas, todėl tikslesniems tyrimams pasirinkta surinkti pilnus naujienų tekstus. Pagal turimas nuorodas į konkrečios naujienos puslapį analizuojami svetainių išeities kodai, taip išgaunant pilną naujienos tekstą. Informacijos surinkimui realizuoti pasitelkiami trečiųjų šalių komponentai:

- „JDOM“ – patikima biblioteka skirta skaityti ir rašyti XML duomenis nenaudojant sudėtingų ir daug atminties užimančių schemų.
- „ROME Fetcher“ – įgalina įrašų parsisiuntimą per HTTP protokolą.
- „jsoup“ – biblioteka leidžianti realiu laiku analizuoti HTML puslapius.

Svarbu paminėti, kad sistemoje saugomos tik naujausios naujienos. Periodiškai, kaip ir informacijos surinkimas, atliekamas saugomų naujienų straipsnių naujumo tikrinamas – ar dienų skirtumas tarp esamos datos ir naujienos publikavimo datos nėra didesnis už nustatytą straipsnių saugojimo datą (pradinė reikšmė 7 dienos). Jei naujienos straipsnis

saugomas ilgiau nei nurodytą laiką – jis ištrinamas iš sistemos. Sistemos prototipe pasirinkta saugoti tik konkretaus laikotarpio naujienas, nes siekiama riboti pradinių duomenų rinkinio dydį, taip susitelkiant tik į aktualiausias naujienas.

4.4.6. Prasminių žodžių išskyrimo realizacija

Kuriant autonomiškas sistemas, vienas iš pasiruošimo darbų yra tinkamai paruošti ir pateikti duomenų rinkinius. Naujienų straipsniai ir jų tematikos yra lengvai suprantamos žmonėms, tačiau tam, kad mašina galėtų sukurti ryšius tarp duomenų rinkinių, klasifikuoti ar sisteminti informaciją, pradiniai duomenys turi būti apdoroti. Pagal atliktą tekstinių duomenų paruošimo analizę (2.5.2 Duomenų paruošimas) sukurta procedūra, kuri atsakinga už naujienų testinės informacijos mažinimą neprarandant svarbios informacijos. Procedūros veikimas susideda iš šių fazių:

1. Analizuojamas naujienų srautas - visi tekstai skaidomi į atskirus žodžius, skaičiuojant jų dažnumą, taip sudaromas bendras žodžių žodynas, kuriame saugomas kiekvieno žodžio pasikartojimo dažnis visame naujienų sraute.

3 lentelė Dažniausių žodžių žodyno pradžia

Žodis	Dažnis	Žodis	Dažnis
kad	9322	dar	2133
yra	5030	tačiau	2112
tai	4275	pat	2064
buvo	4218	bus	2008
kaip	3261	jau	1938
apie	3031	jis	1906
taip	3016	prieš	1905
savo	2956	kai	1897
lietuvos	2928	iki	1849
tik	2806	bei	1798
bet	2737	gali	1745
dėl	2570	daugiau	1560
nuo	2417	prie	1527
per	2317	metu	1521
metų	2151	kas	1397

2. Pagal gautą žodžių žodyną, iškeliant prielaidą, kad apie 1-2 procentai visame naujienų sraute naudojamo žodyno sudaro dažniausi žodžiai - išvedama formuluotę, pagal kurią sistema atrenka nereikšminius žodžius iš sukurto žodyno. Pagal turimą

dažniausių žodžių sąrašą apdorojami visi naujienų tekstai išrenkant nereikšminius žodžius.

3. Realizuojamas „žodžių maišo“ (angl. „bag-of-words“) modelis – naujienų atskiriems likusiems žodžiams skaičiuojami pasikartojimo dažnumai.

Vykdymo pabaigoje pagal gautus rezultatus galime kiekvienai naujienai išskirti prasminius žodžius. Prasminiai naujienos žodžiai yra tie, kurių pasikartojimo dažnis yra ženkliai didesnis už kitų žodžių dažnį. Atitinkamai prasminiai naujienos žodžiai duomenų bazėje saugomi kai jų pasikartojimų dažnis naujienos tekste yra didesnis už vienetą.

4.4.7. Naujienų tekstų kamienizavimo realizacija

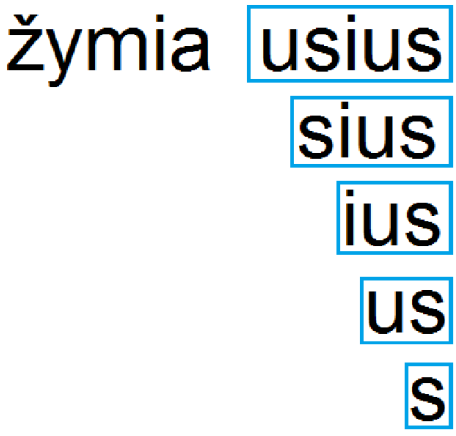
Kompiuterinės sistemos tekstinę informaciją lygina pagal visą simbolių seką, todėl atsiranda poreikis sukurti algoritmą, kuris teiktu galimybę sujungti skirtingas galūnes, bet tą pačią prasmę turinčius žodžius. Vienas iš siūlomų metodų – žodžių kamienizavimas (2.5.2 Duomenų paruošimas). Populiarioms kalboms yra sukurti atskiri kamienizavimo įrankiai (2.7 Įrankių analizė), kurių pagrindinis veikimo principas yra nustatyti konkrečiai kalbai būdingą žodžių galūnių sąrašą, pagal kurį analizuojami norimi kamienizuoti žodžiai. Kadangi kuriama autonominė sistema, pasirinktas ir realizuotas kamienizavimo algoritmas, kuris nenaudoja iš anksto nustatytų statinių reikšmių (šiuo atveju Lietuvių kalbai būdingų žodžių galūnių). Analogiškai prasminių žodžių iškyrimo realizacijai kurtas kamienizavimo algoritmas susideda iš šių fazių:

1. randamos dažniausios galūnės visame naujienų sraute,
2. kamienizuojami viso naujienų srauto naujienos pagal rastas dažniausias galūnes.

Dažniausių galūnių radimas pavaizduotas veiklos diagramoje, kuri pateikiama žemiau.

7 pav. Dažniausių žodžių galūnių radimo veiklos diagrama

Žodžių galūnių radimas vykdomas kai visi naujienų srauto tekstų žodžiai analizuojami po vieną. Konkretaus žodžio galūnės analizė atliekama taip: žodis dalinamas į dvi dalis, antrosios žodžio dalies raidžių deriniams skaičiuojami pasikartojimo dažniai visame naujienų žodžių sraute. Žodžio galūnės derinių sudarymo pavyzdys pateikiamas 6 paveikslėlyje.



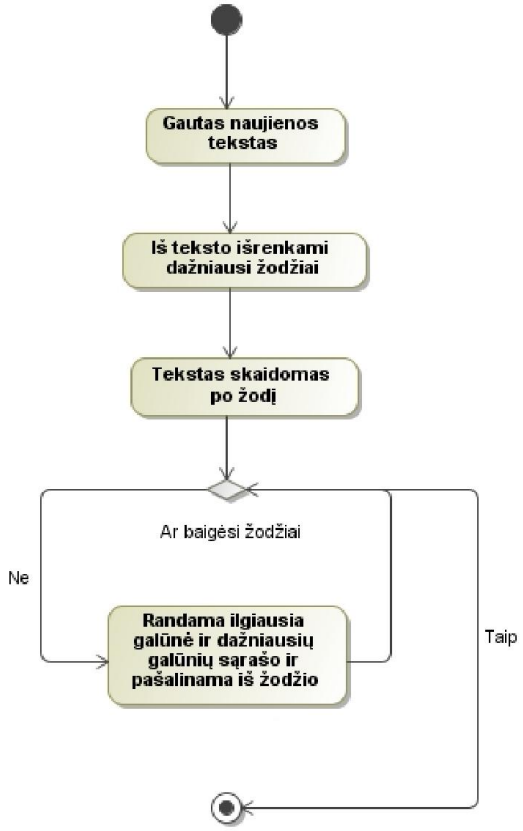
8 pav. Žodžio galūnės derinių skaidymas

Rezultate gaunamas galūnių ir jų dažnių sąrašas (4 lentelė).

lentelė 4 Dažniausių žodžių galūnių sąrašas

Žodžio galūnė	Dažnis	Žodžio galūnė	Dažnis
s	74960	mo	9916
i	59226	iai	9232
o	49541	ja	9096
ų	44239	į	8897
a	41552	mas	8853
as	36295	jos	8634
e	35878	t	8322
os	30192	ijos	8316
ai	25799	ius	8113
ą	25635	tas	7810
is	25272	oje	7647
ti	24346	ms	7314
us	18507	me	7183
ių	18121	tų	6896
ė	16270	jo	6868
u	15659	ę	6637
ės	13389	ia	6504
je	12059	mą	6491
io	10322	jų	6378

Toliau, pagal gautą dažniausių galūnių sąrašą kamienizuojami visi naujienų teksto žodžiai. Analizės eiga pateikta žemiau.



9 pav. Naujienos teksto kamienizavimo veiklos diagrama

Kamienizavimas konkrečiam žodžiui atliekamas taip: iš dažniausių galūnių sąrašo ieškoma ilgiausia galūnė iš sudaryto galūnių sąrašo, su didžiausių dažniu, kuri bus atmetama iš nagrinėjamo žodžio pabaigos. Atliekant žodžio kamienizavimą laikomasi apribojimo kad nuo žodžio galo ištrinamas raidžių junginys negali viršyti pusės žodžio ilgio.

Kamienizavimo pavyzdys: kamienizuojamas žodis „žymiausius“. Iš dažniausių galūnių sąrašo išrenkamos nagrinėjamam žodžiui tinkančios galūnės ir jų dažniai: „s“- 74960, „us“- 18507, „ius“-8113, „sius“-878 (pagal 4 lentelės duomenis). Iš nagrinėjamo žodžio pabaigos ištrinama ilgiausia galūnė su didžiausiu dažniu - ilgiausia galūnė yra „sius“, kitos galūnės yra trumpesnės todėl netinkamos. Kamienizavimo rezultatas - „žymiau“. Kamienizavimo rezultatai priklauso nuo sudaryto dažniausių galūnių sąrašo ir norimo kamienizuoti žodžio ilgumo.

Svarbu paminėti, kad kamienizuojamas naujienos tekstas, kuriame pašalinti nereikšminiai žodžiai, nes dažnai tekste randamų jungtukų ar kitų kalbos dalių

kamienizuoti neįmanoma. Gautas kamienizuotas naujienos tekstas saugomas duomenų bazėje, taip pat prasminiams žodžių kamienams saugoti realizuotas „žodžių maišo“ (angl. „bag-of-words“) modelis - prasminiai kamienizuoto teksto unikalūs žodžiai saugomi kartu su jų pasikartojimo dažnumais.

4.4.8. Naujienų grupavimo realizacija

Naujienų grupavimui pagal atliktą analizę (žiūrėti skyriuje 2.7 Įrankių analizė), pasirinktas „carrot^2“ įrankis. Šis įrankis pasirinktas, nes yra plačiai naudojamas kituose mokslinio pobūdžio tyrimuose bei pateikia tris skirtingus algoritmus tekstinės informacijos grupavimui.

Naujienų grupavimui naudojami „STC“, „Lingo“ ir „Bisecting K-means“ algoritmai (plačiau apie šiuos algoritmus žiūrėti skyriuje 2.7.1 Įrankis „carrot2“).

4.4.9. Naujienų atvaizdavimo realizacija

Informacijos atvaizdavimui pasirinkta naudoti „PrimeFace“ - internetinių puslapių kūrimo komponentų biblioteka. Sukurtame sistemos prototipe atvaizduojama tokia informacija:

- Grafinė informacija: naujienų pasiskirstymas pagal šaltinius ir dienas.
- Dažniausių raktažodžių (prasminių žodžių) ir kamienizuotų raktažodžių sąrašas.
- Pradinis kategorijų ir naujienų sąrašas, kuris gautas iš naujienų šaltinių.
- Naujienų grupių sąrašas ir joms priskirtos naujienos.

Sistemos vaizdai pateikti priede (0 skyrius).

5. EKSPERIMENTINĖ DALIS

Eksperimentu siekiama ištirti kaip kinta naujienų grupavimo rezultatai taikant skirtingas naujienų tekstų mažinamo metodikas. Matuojami gautų naujienų kategorijų kiekiai, bei joms priskirtos naujienos. Analizuojama ar kategorijai priskirtos naujienos atspindi kategorijos sugeneruotą pavadinimą (kokie yra kategorijos pavadinimo ryšiai su priskirtų naujienų tekstynu), rezultatai lyginami naudojant skirtingus grupavimo algoritmus.

Eksperimentams atlikti naudojamų duomenų atrinkimo kriterijai:

- Atrenkamas nustatyto dienų intervalo naujienų sąrašas, kuris nesikeis viso tyrimo metu.
- Naudojami naujienų tekstai, kurių pradinis žodžių skaičius didesnis negu 300 žodžių. Taip išvengiama netikslumų ir rezultatų iškraipymų, kurie galimi jei lygintume tekstus, kurių dydžiai kardinaliai skiriasi.

Pagal šiuos kriterijus atrenkami pradiniai duomenys. Viso per 11 dienų surinkta ir eksperimentams atlikti naudojama 902 naujienų.

5.1. Naujienų tekstų apimties mažinimo rezultatai

Tiriama, kaip kinta naujieną reprezentuojantis tekstas, taikant skirtingus mažinimo metodus. Išskiriamos dvi pagrindinės naujienų tekstų mažinimo metodikos: prasminių žodžių išskyrimo ir žodžių kamienizavimo. Svarbu paminėti, kad prieš prasminių žodžių iškyrimą ir kamienizavimą atliekamas nereikšminių žodžių šalinimas. Kaip informacijos mažinimas įtakoją naujienų grupavimą nagrinėjama skyriuje 5.2 „Naujienų grupavimo eksperimentas“.

5.1.1. Prasminių žodžių išskyrimo rezultatai

Prasminių žodžių išskyrimo realizacija aprašyta skyriuje 4.4.6 „Prasminių žodžių išskyrimo realizacija“. Pagal realizuotą algoritmą gauti tokie rezultatai:

- Sudarytas bendras viso naujienų srauto žodžių žodynas, kurį sudaro virš 95 tūkstančiai žodžių.
- Sistemai renkantis, kurie žodžiai bus traktuojami kaip nereikšminiai - iš bendro žodyno sudaromas sąrašas žodžių ir jų dažnių, kuris rūšiuojamas pagal dažnį, mažėjančiai. Stebima žodžių dažnių seka, ties pastebimai staigesniu žodžių

dažnių sumažėjimu dedama riba, atskirianti, kurie žodžiai bus nereikšminiai. Vidutiniškai gaunama dviejų procentų riba (viso virš 3000 žodžių).

- Pradiniuose naujienų tekstuose vidutiniškai yra 583,21 žodžių. Iš pradinių naujienų tekstų išėmus nereikšminius žodžius gautas vidutis naujienų tekstų žodžių skaičius yra 403,66 tai reiškia, kad tekstų dydžiai sumažėjo31procentais.
- Išskiriant prasminius naujienos žodžius, kurie naujienos tekste pasikartojo daugiau nei vieną kartą, gauname, kad vidutiniškai naujienos turi po 49,53prasminių žodžių.

Apibendrinus galima teigti, kad vidutiniškai trečdalis naujienos teksto yra nereikšminiai žodžiai, kurie nėra informatyvūs. Prasminių naujienos žodžių iškyrimas gali sumažinti naujienų tekstinę informaciją 91,5 procentais.

5.1.2. Žodžių kamienizavimo rezultatai

Žodžių kamienizavimo algoritmo realizacija aprašyta skyriuje 4.4.7 „Naujienų tekstų kamienizavimo realizacija“. Pagal realizuotą algoritmą gauti tokie rezultatai:

- Sudarytas viso naujienų srauto galūnių sąrašas, kurį sudaro 2050 galūnių (sąrašo pradžia pateikta 4 lentelėje).
- Išskyrus prasmines kamienizuotas žodžio dalis tekstuose gaunama, kad vidutiniškai naujieną reprezentuoja 57,89 kamienizuoti prasminiai žodžiai.
- Kamienizavus naujienos tekstą ir išrinkus prasminius kamienus (prasminiai kamienai laikomi, tada kai naujienos tekste pasikartojo daugiau nei vieną kartą), vidutiniškai teksto dydis sumažėja per 90,07 procentų. Primename, kad kamienizuojamas tekstas, kuriame pašalinti nereikšminiai žodžiai.

5.1.3. Tekstų mažinimo rezultatų apibendrinimas

Visas naujienų srautas sudarytas iš 902 naujienų, kurių bendras žodžių skaičius yra 526 061. Išskyrus prasminius naujienų žodžius gautas bendras naujienų srauto prasminių žodžių skaičius 364 103. Atrinkus prasminius kamienizuotus naujienų tekstus gautas bendras naujienų srauto prasminių kamienizuotų žodžių skaičius yra 52 217. Procentiniai didžiausias sumažėjimas gaunamas atlikus prasminių žodžių išskyrimą, naujienos žodžių skaičius sumažėja per 91,5 procentus lyginant su originaliu žodžių skaičiumi.

Taip pat, pagal realizuotus teksto mažinimo metodus atlikta analizė, kaip kinta skirtingų ilgių naujienų tekstų dydžiai, pateikiama naujienų tekstų dydžių kaitos lentelė (5 lentelė). Lentelėje pateikiami naujienų tekstų vidutiniai dydžiai atitinkamai pagal pradinių

tekstų dydžių intervalus. Taip pat skaičiuojama, kiek procentų sumažėja naujienos žodžių skaičius lyginant su pradiniu kiekiu.

5 lentelė Naujienų dydžių kitimo statistika

Pradinių tekstų dydžių intervalai	**[300-700]**	**[701-1200]**	**[1201-1700]**	**[1701-2700]**
Naujienų skaičius	673	187	31	11
Vidutinis pradinių žodžių skaičius	443.73	872.29	1356.9	2022.18
Vidutinis žodžių skaičius atmetus nereikšminius žodžius	308.9 (31.26%)	595.21 (31.75%)	951.67 (29.86%)	1400.18 (30.75%)
Vidutinis prasminių žodžių skaičius	35.08 (92.2%)	78.82 (90.9%)	131.58 (90.30%)	204.36 (89.89%)
Vidutinis kamienizuotų prasminių žodžių skaičius	41.67 (90.5%)	90.85 (89.5%)	150.9 (88.87%)	227.54 (88.74%)

Pagal rezultatų pasiskirstymą matome, kad daugiausiai naujienų, kurių pradiniai tekstai turi nuo 300 iki 700 žodžių. Išskiriant prasminius žodžius naujienų žodžių skaičių galima sumažinti iki 90 procentų lyginant su pradiniu žodžių skaičiumi. Taip pat matome, kad žodžių mažinimo rezultatai, taikant skirtingus metodus mažai svyruoja tarp skirtingų dydžių naujienų, todėl galima teigti, kad nereikšminių ir prasminių žodžių kiekiai proporcingi naujienų dydžiams.

5.2. Naujienų grupavimo eksperimentas

Eksperimentu siekiama ištirti kaip kinta naujienų grupavimo rezultatai taikant skirtingas naujienų tekstų mažinamo metodikas. Matuojami gautų kategorijų kiekiai, bei joms priskirtos naujienos ir analizuojami priskirti kategorijų pavadinimai. Analizuojama ar grupei priskirtos naujienos atspindi kategorijai sugeneruotą pavadinimą. Rezultatai lyginami naudojant skirtingus grupavimo algoritmus ir paduodamų duomenų rinkinius.

Pagrindinis eksperimento tikslas yra išsiaiškinti, kokią taktiką naudojant galima išgauti konkrečiausias kategorijas ir prasmingiausius naujienų kategorijų pavadinimus.

5.2.1. Eksperimentiniai duomenų rinkiniai ir vertinimo kriterijai

Tyrime naudojami duomenų rinkiniai, sudaryti pagal skyriuose 4.4.6 „Prasminių žodžių išskyrimo realizacija“ ir 4.4.7 „Naujienų tekstų kamienizavimo realizacija“ pateiktą aprašymą. Sudaryti tokie naujienų tekstų rinkiniai:

1. Pradinis ir originalus naujienos tekstas.
2. Tekstas, kuriame pašalinti nereikšminiai žodžiai (žodžiai gali kartotis bei išlaikomas žodžių eiliškumas).
3. Tekstas, sudarytas tik iš unikalių naujienos prasminių žodžių.
4. Kamienizuotas tekstas, kuriame pašalinti nereikšminiai žodžiai (kamienizuoti žodžiai gali kartotis, išlaikomas žodžių eiliškumas).
5. Kamienizuotas tekstas, sudarytas tik iš unikalių naujienos prasminių kamienizuotų žodžių.

Naujienų tekstų grupavimui, naudojamas „carrot^{2}“ įrankis, kuris pateikia trijų algoritmų realizacijas: „Lingo“, „STC“, „Bisecting K-means“ (plačiau žiūrėti skyriuje 2.7.1 Įrankis „carrot2“).

Naujienų kategorijų pavadinimų prasmingumas ir tinkamumas matuojamas pagal tai, kokie žodžiai sudaro sugeneruotą pavadinimą ir kokia šių žodžių reikšmė bendrame naujienų sraute. Naujienų kategorijų pavadinimo prasmingumas įvertinamas taip:

1. Kategorijos pavadinimas skaidomas į atskirus žodžius.
2. Kiekvienam žodžiui surandamas atitikmuo ir jo pasikartojimo dažnis bendrame naujienų srauto žodyne (žiūrėti skyriuje 4.4.6 „Prasminių žodžių išskyrimo realizacija“), jei analizuojami kamienizuoti kategorijos žodžiai, tai atitikmenys nustatomi ieškant žodyne žodžių, kurie turi analizuojamą kamieną ir imamams maksimalų dažnį turintis žodis.
3. Kategorijai skaičiuojami surastų atitikmenų dažnių vidurkiai.

Pateikiamas kategorijos pavadinimo įvertinimo skaičiavimo pavyzdys: skaičiuojamas įvertinimas kategorijai, kurios sugeneruotas pavadinimas yra „Sveikatos, Gyvenimo, Gyventojų“. Pirma randami kiekvieno žodžio dažniai iš bendro žodžių žodyno - „Sveikatos - 272, Gyvenimo - 275, Gyventojų - 295“. Apskaičiuojamas bendras kategorijos vidurkis yra 280, taigi šios kategorijos įvertis būtų 280.

Skaičiuojamas kategorijos žodžių dažnis parodo, kokia šios kategorijos reikšmė visame naujienų sraute. Kategorijos, kurių įverčiai didžiausi ir pavadinime yra neprasminių žodžių, laikomos bendrinėmis ir neteikiančiomis naudos.

Turint visų kategorijų pavadinimų įvertinimus galima apskaičiuoti bendrą grupavimo įvertinimą išvedant kategorijų įvertinimų vidurkį. Daroma prielaida, kad kuo apskaičiuotas įvertis mažesnis tuo sudarytų kategorijų pavadinimai yra tikslesni ir kategorijų pavadinimuose nepasitaiko nereikšminių žodžių.

5.2.2. Naujienų grupavimo tyrimas

Atitinkamai, gauti 902 naujienų grupavimo rezultatai skirstomi pagal algoritmus ir naudotus duomenų rinkinius (6 lentelė). Rezultatų lentelėje duomenų rinkinių pavadinimai trumpinami taip:

- T1 - Originalūs naujienų tekstai;
- T2 - Tekstai, kuriuose pašalinti nereikšminiai žodžiai;
- T3 - Kamienizuoti tekstai;
- T4 - Tekstai, kurie sudaryti iš prasminių unikaliu žodžių;
- T5 - Kamienizuoti tekstai, kurie sudaryti tik iš prasminių unikalių kamienizuotų žodžių.

6 lentelė Naujienų grupavimo įverčiai gauti naudojant skirtingus duomenų rinkinius ir algoritmus

Algoritmai	Duomenų rinkiniai	Bendras grupavimo įvertis	Naujienų kategorijų skaičius	Vidutinis naujienų skaičius kategorijoje
„Lingo“	T1	1190,21	83	93,60
	T2	198,53	88	25,69
	T3	282,94	89	60,29
	T4	230,37	80	25,48
	T5	282,01	87	18,24
„STC“	T1	1305,37	16	454,06
	T2	242,55	16	154,25
	T3	1456,73	16	276,81
	T4	268,26	16	71,87
	T5	869,86	16	115,75
„Bisecting K-means“	T1	1300,82	25	36,08
	T2	262,84	25	730,00
	T3	9322,00	25	113,96
	T4	265,33	25	137,92
	T5	9322,00	25	808,16

Pagal gautus rezultatus matomas aiškus skirtumas tarp naudojamų algoritmų ir generuojamų kategorijų skaičiaus. Daugiausiai automatiškai generuotų kategorijų gauta naudojantis „Lingo“ algoritmą, nes šis algoritmas grupavimui naudoja dažniausių panašių

frazių paiešką tekste. Tekstuose išskyrus daug skirtingų frazių (frazės gali būti ir pavieniai žodžiai), gauname didesnį kiekį grupių, kuriose yra santykinai mažai naujienų.

Analizuojant grupavimą pagal duomenų rinkinius gauname, kad naujienų grupių įvertis yra didžiausiais atliekant originalių arba kamienizuotų tekstų grupavimą. Tokie rezultatai byloja, kad kategorijų, kurios gautos iš originalaus arba kamienizuoto teksto, pavadinimai yra neaiškūs ar bendriniai.

Grupavimo įverčiai yra mažiausi atliekant grupavimą su duomenų rinkiniais, kai tekstuose pašalinti nereikšminiai žodžiai ir kai tekstai sudaryti tik iš prasminių žodžių. Detaliau išanalizavus gautus grupavimo rezultatus pastebėta, kad su tekstais, kuriuose pašalinti tik nereikšminiai žodžiai, gautos kategorijos sudarytos iš didesnio kiekio žodžių, nei kategorijos gautos atlikus grupavimą su tekstais, sudarytais iš prasminių žodžių, todėl galima teigti, kad prasmingiausios kategorijos gaunamos, kai grupavimui naudojamas tekstas sudarytas tik iš prasminių žodžių. Keletas kategorijų pavadinimų pavydžių atliekant grupavimą su tekstais be nereikšminių žodžių: „Automobilio, Pinigų, Savivaldybės“, „Rūšis Standartinė Publikacija“, „Gyvenimo, Pinigų, Žmogaus“. Kai grupavimui naudojami tekstai sudaryti tik iš prasminių žodžių, daugumą sugeneruotų kategorijų susideda iš vieno žodžio, taip pat sudarytose kategorijose yra mažesnis naujienų kiekis lyginant su rezultatais gautais grupuojant tekstus be nereikšminių žodžių. Automatiškai sugeneruotų naujienų grupių detalesnė informacija pateikta prieduose.

5.3. Eksperimento apibendrinimas

- Atlikus naujienų teksto mažinimą, išmetant nereikšminius žodžius, nustatytą, kad 30,78 procento originalaus teksto sudaro nereikšminiai žodžiai. Taip pat įrodyta, kad nereikšminių ir prasminių žodžių kiekiai yra proporcingi naujienų tekstų dydžiams, nes teksto mažinimo procentinės reikšmės mažai skiriasi lyginant įvairių dydžių tekstus (5 lentelė).
- Žodžių kamienizavimas tinkamas kai norima sujungti tą pačią prasmę, bet skirtingas galūnes turinčius žodžius. Pagal duomenis, kurie pateikti 5 lentelėje, matome, kad prasminių kamienizuotų žodžių naujienoje išgaunama daugiau negu nekeistų prasminių žodžių.
- Grupavimas laikomas tikslus, kai sugeneruotų naujienų grupių ir naujienų skaičiaus grupėje pasiskirstymai yra proporcingi, todėl stabiliausias naujienų grupavimas gautas naudojantis „Lingo“ algoritmu.
- Grupavimas atliktas naudojantis originaliais ir kamienizuotais tekstais įvertintas blogiausiai, nes naujienų kategorijos sudarytos iš per daug bendrinių ir neaiškių žodžių.
- Prasmingiausios kategorijos, pagal apskaičiuotą įvertį ir vidutinį naujienų kiekį grupėje, gautos naudojantis naujienų tekstais, kurie sudaryti iš unikalių prasminių žodžių.

6. ĮŽVALGOS IR TOLESNĖS SISTEMOS PLĖTOJIMO GALIMYBĖS

Šiame skyriuje pateikiama informacija apie duomenis ir jų ryšius, kurie nepateko į pagrindinį tyrimą, bei numatomas sistemos plėtojimo galimybes ir siūlymus. Sistemos prototipo realizacijos metu ir atliekant tyrimą buvo pastebėti gaunamų duomenų sąryšiai, kurie gali būti naudingi tolesnei sistemos plėtrai. Taip pat šiame skyriuje pateikiami papildomi statistiniai duomenys.

6.1. Raktažodžių kaita laike ir naujienų istorijų sekimas

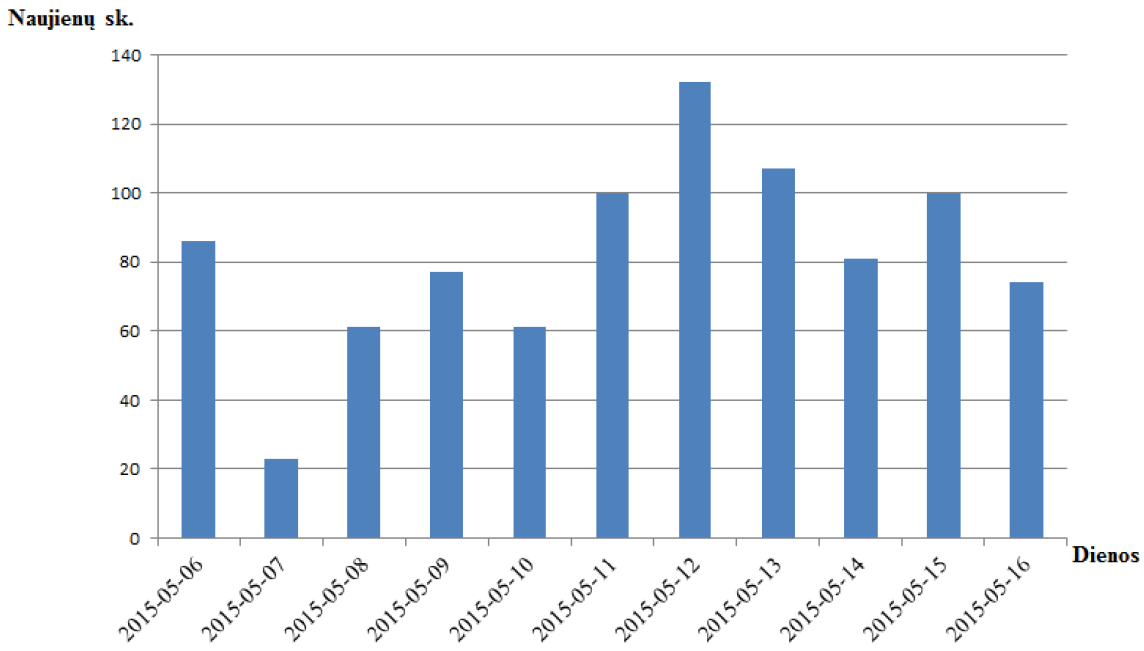
Naujienų prasminius žodžius galima laikyti raktažodžiais, pagal kuriuos galima atliktį įvairią naujienų srauto analizę, filtravimą ar susijusių naujienų paiešką.

Atliekant naujienų tekstų prasminių žodžių ir jų kamienų analizę išsaugomi visi prasminiai žodžiai(raktažodžiai) kartu su jų pasikartojimo dažniu, todėl galima išrinkti populiariausius naujienų srauto raktažodžius. Pagal tyrime naudotus duomenis dažniausių raktažodžių sąrašas pateiktas 7 lentelėje.

7 lentelė Populiariausi prasminiai žodžiai

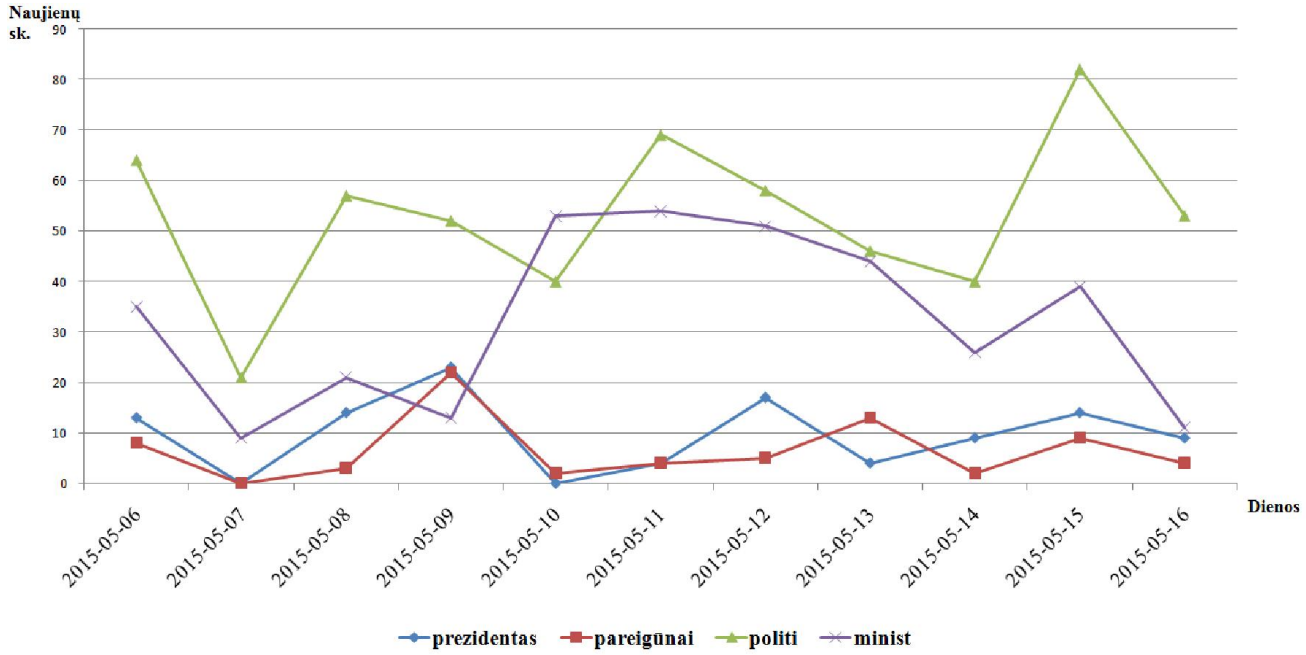
Raktažodžiai	Susijusių naujienų skaičius	Kamienizuoti raktažodžiai	Susijusių naujienų skaičius
prezidentas	64	toki	212
vakarų	64	politi	175
komandos	61	keli	158
komanda	59	galimy	154
sąjungos	59	nauj	153
vyras	58	tyrim	151
pareigūnai	58	gyveni	133
klaipėdos	57	kuria	133
sistemos	57	minist	131
ukrainoje	57	klausi	128

Naujienų pasiskirstymas pagal dienas (10 pav.) gali suteikti tik kiekybinę informaciją: kiek dienų kaupiamos/saugomos naujienos, kiek naujienų publikuojama ir kita.



10 pav. Naujienų pasiskirstymas pagal dienas

Vertingiau yra analizuoti kaip raktažodžiai pasiskirsto laike. Sudaroma diagrama, kurioje pateikiama informacija, kaip raktažodžių populiarumas (susijusių naujienų skaičius) kinta laike (12 pav.).



11 pav. Populiariausių raktažodžių pasiskirstymas laike

Analizei išsirinkti raktažodžiai („prezidentas“, „pareigūnai“) ir kamienizuoti raktažodžiai („politi“, „minist“). Pagal raktažodžių pasiskirstymą laike galime matyti, kad proporcingai kinta raktažodžių „politi“ ir „minist“ kreivės. Šių raktažodžių kitimas laike panašus, nes kreivių augimo ir leidimosi pradžios taškai sutampa (2015-05-07, 2015-05-08, 2015-05-14, 2015-05-15) bei proporcingai kinta ir susijusių naujienų skaičiai. Toks kreivių panašumas gali reikšti, kad egzistuoja naujienų grupė, kurioje nagrinėjamos panašios tematikos aktualijos.

Kita įdomi duomenų koreliacija matoma tarp raktažodžių „prezidentas“ ir „pareigūnai“. Šiems raktažodžių kreivėje, taške 2015-05-09 sutampa naujienų skaičius, bei šios kreivės identiškai auga nuo 2015-05-07 iki 2015-05-09 ir mažėja nuo 2015-05-08 iki 2015-05-10. Tokie rezultatai gali byloti apie naujienų sraute iškilusią konkrečią naujienų tematika, kuri yra aktuali tik ribotą laiko kiekį.

Nagrinėjant raktažodžių pasiskirstymą laike galima daryti šias prielaidas: išskiriant proporcingai kintančias kreives galima grupuoti naujienas, bei, ribotą laiko tarpą sutampančios kreivės, gali išskirti konkrečias populiariausias aktualijas. Tačiau šioms prielaidoms patvirtinti reikalingi tikslesni tyrimai.

6.2. Sistemos plėtojimo perspektyvos

Šiuo metu sistemoje pateikiama informacija, kuri yra surenkama iš konkretaus naujienų šaltinių sąrašo. Šį sąrašą gali redaguoti tik sistemos administratorius. Viena iš prototipo tobulinimo krypčių būtų sistemos personalizavimas. Pagrindinis funkcionalumas leistu prisijungti skirtingiems vartotojams, kurie galėtų pritaikyti sistemą pagal konkrečius poreikius (keisti šaltinių sąrašą, filtruoti naujienų srautą ir kita).

Kita sistemos plėtros kryptis būtų gerinti naujienų grupavimą realizuojant naujienų klasifikavimą. Naujienų klasifikavimo esmė yra sukurti unikalias klases, kurioms būtų priskiriamos naujienos. Klasių formavimui būtų galima panaudoti išrinktus prasminius naujienų žodžius. Siūlomas naujienų klasių formavimo algoritmas susidėtų iš šių pagrindinių veiksmų:

1. Iš pradinių duomenų, kurie gaunami tiesiai iš naujienų šaltinių, išsirenkama aibė naujienų. Išsirinktos naujienos turi turėti pradines autorių priskirtas prasmines kategorijas (pvz. „Politika“, „Kriminalai“).
2. Išsirinktai aibei naujienų vykdomas prasminių žodžių išskyrimas (4.4.6 „Prasminių žodžių išskyrimo realizacija“).

3. Išsirenkama viena pradinė iš naujienų šaltinių gauta kategorija, kuriai priskiriami visi gauti prasminiai žodžiai iš priskirtų naujienų.

Tikėtinas naujienų klasės pavyzdys būtu toks: klasės pavadinimas - „Sportas“, klasei priskirti prasminiai žodžiai: „krepšinis“, „futbolas“, „rezultatas“ ir kiti. Naujienų klasifikavimo pagrindinis išskirtinumas – nenaudojami statiniai duomenų rinkiniai, klasių sudarymas vyktų dinamiškai, su kiekviena naujų duomenų aibe sistema „mokinasi“ atskirti naujienų klases. Norint klasifikuoti konkretų naujienų sąrašą reikėtų tik ieškoti, kurios klasės priskirti žodžiai geriausiai atitinka nagrinėjamos naujienos prasminius žodžius.

7. IŠVADOS

- Eksperimento metu buvo siekiama išsiaiškinti su kokiais naujienų tekstų rinkiniais gaunamos prasmingiausios kategorijos. Pagal gautus rezultatus (žiūrėti 6 lentelė) nustatytą, kad prasmingiausios kategorijos gaunamos atliekant grupavimą su naujienomis, kurių tekstai sudaryti tik iš prasminių unikalių žodžių.
- Grupavimas laikomas tikslus, kai sugeneruotų naujienų grupių ir naujienų skaičiaus grupėje pasiskirstymai yra proporcingi (žiūrėti 6 lentelė), todėl stabiliausias naujienų grupavimas gautas naudojantis „Lingo“ algoritmu.
- Tai pat pastebėta, kad su tekstais, kuriuose pašalinti tik nereikšminiai žodžiai, gautos kategorijos sudarytos iš didesnio kiekio žodžių, nei kategorijos gautos atlikus grupavimą su tekstais, sudarytais iš prasminių žodžių.
- Nagrinėjant raktažodžių pasiskirstymą laike (žiūrėti 12 pav.) galima daryti šias prielaidas: ieškant proporcingai kintančių kreivių galima išskirti naujienų grupes. Taip pat ribotą laiko tarpą sutampančios kreivės, gali byloti apie konkrečių populiariausių aktualijų iškilimą.
- Žodžių kamienizavimas tinkamas kai norima sujungti tą pačią prasmę, bet skirtingas galūnes turinčius žodžius. Pagal duomenis, kurie pateikti 5 lentelėje, matome, kad prasminių kamienizuotų žodžių naujienoje išgaunama daugiau negu nekeistų prasminių žodžių. Tačiau naujienų grupavimo rezultatai parodė, kad naudojant kamienizuotus naujienos žodžius gaunamos abstrakčios kategorijos, todėl žodžių kamienizavimas labiau tinkamas naujienų filtravimui arba raktinių žodžių analizei.

8. LITERATŪROS SĄRAŠAS

[1] Isbell K., „The Rise of the News Aggregator: Legal Implications and Best Practices“ American Society of Clinical Oncology (ASCO); Berkman Centerfor Internet &Society, Berkman Center Research Publication, 2010

[2] Simec, A.; Carapina, M.; Duk, S., „RSS as medium for information and communication technology“, MIPRO, 2011 Proceeding soft he 34th International Convention , pp.1593,1596, 23-27 May 2011

[3] Chowdhury S., Landoni M., „News aggregator services: user expectations and experience“ , Online Information Review, Vol. 30 Iss: 2, pp.100 – 115, 2006

[4] „Top 5 Business IT Trends to 2020” [žiūrėta 2013-11-16]. Prieiga per internetą:https://www.portal.euromonitor.com/Portal/Pages/Common/Pdf.aspx/Top_5_Business_IT_Trends_to_2020

[5] „Statistika. Internetu naudojasi 69 proc. Lietuvos gyventojų“ [žiūrėta 2013-10-03]. Prieiga per internetą: http://www.ivpk.lt/news/1764/61/Internetu-naudojasi-69-proc-Lietuvos-gyventoju.html

[6] Autorių teisių ir gretutinių teisių įstatymas [žiūrėta 2013-11-15]. Prieiga per internetą:http://www3.lrs.lt/pls/inter2/dokpaieska.showdoc_l?p_id=207199

[7] Han Y. G; Lee S.H.; Kim J.H.; Kim Y., „A New Aggregation Policy for RSS Services“, Proceedings of the 2008 internation alworkshop on Context enabled source and services election, integration and adaptation: organized with the 17th International World Wide Web Conference, ArticleNo. 2 , 2008

[8] Bharath S., „Short text classifikation in Twitter to improve information filtering“, The Ohio State University 2010

[9] Bock H.-H.; „Clustering Methods: A History of K-Means Algorithms“, Institute of Statistics, RWTH Aachen University, D-52056 Aachen, Germany, 2007

[10] Xie J.; Jiang S., "A Simple and Fast Algorithm for Global K-means Clustering, Education Technology and Computer Science (ETCS), 2010 Second International Workshopon , pp. 36,40, 6-7 March, 2010

[11] Steinbach M.; Karypis G.;Kumar, V. „A Comparison of Document Clustering Techniques” , Technical report, University of Minnesota, 2000

[12] Bradleyand P. S. ;Fayyad U. M, „Refining initial points for k-means clustering” Proceeding Fifteenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, pp. 91–99, 1998

[13] Sharma, A.; Dhir, R., „A Word sets based document clustering algorithm for large datasets“, Methods and Models in Computer Science, 2009. ICM2CS 2009
Proceeding of International Conferenceon , pp.1,7, 14-15 December, 2009

[14] Pouliquen B.; Steinberger R.; Deguernel O.. „Story tracking: linking similar news over time and across languages". In *Proceeding softhe Workshopon Multi-source Multilingual Information Extraction and Summarization* (MMIES '08) Association for Computational Linguistics, Stroudsburg, PA, USA,pp. 49-56, 2008

[15] Burkepile, A.; Fizzano, P., „Classifying RSS Feeds with an Artificial Immune System", Information, Process, and Knowledge Management, 2010. eKNOW '10, Second International Conference , pp.43,47, 10-15 February, 2010

[16] Kaur U. „Comparis on Between K-Mean and Hierarchical Algorithm Using Query Redirection", International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), pp. 1454-1459, June 2013

[17] „Carrot^2" [žiūrėta 2015-05-15]. Prieiga per internetą: http://project.carrot2.org/

[18] „Carrot^2 publications" [žiūrėta 2015-05-15]. Prieiga per internetą: http://project.carrot2.org/publications.html

[19] Osinski S.; Weiss D., „A Concept-Driven Algorithm for Clustering Search Results", Institude of Computing Science , Poznan University of Technology, 2005

[20] Osinski S.; Weiss D., „Conceptual Clusterinf Using Lingo Algorithm: Evaluation on Open Directory Project Data", Institude of Computing Science, Poznan University of Technology, 2005

[21] „ Alias-iCompany" [žiūrėta 2015-05-05]. Prieiga per internetą: http://alias-i.com/lingpipe/index.html

[22] „Mahout-apache" [žiūrėta 2015-04-25]. Prieiga per internetą: http://mahout.apache.org/

[23] Breiman L., „Random forest classifier for remote sensing classification",32-45, February, 2007

[24] „Weka 3: Data MiningSoftwarein Java", [žiūrėta 2015-04-15]. Prieiga per internetą: http://www.cs.waikato.ac.nz/ml/weka/index.html

[25] Sharma N.; Bajpai A.; Litoriya R., „Comparison the various clustering algorithms of weka tools", International Journal of Emerging Technology and Advanced Engineering, Volume2, Issue 5, May 2012

[26] „Snowball, DrMartinPorter", [žiūrėta 2015-02-20]. Prieiga per internetą: http://snowball.tartarus.org/index.php

[27] „ltstemmer", [žiūrėta 2015-04-29]. Prieiga per internetą: http://sourceforge.net/projects/ltstemmer/].php

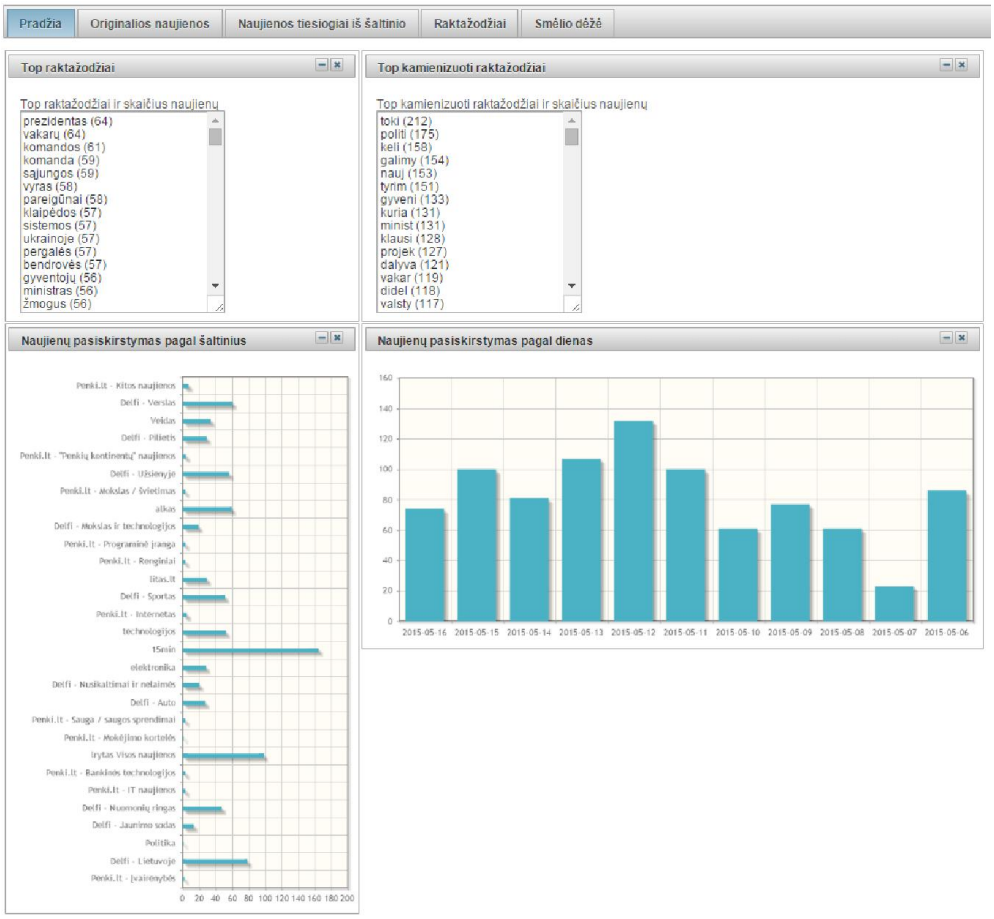
9. PRIEDAI

Naujienų šaltinių sąrašas

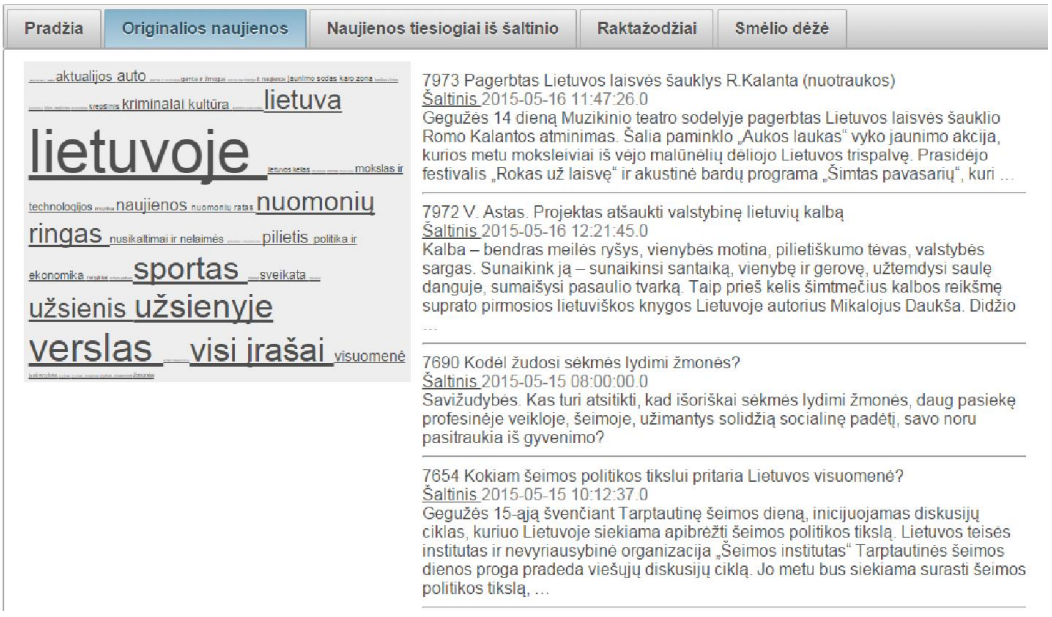
8 lentelė Pagrindinių naujienų srautų adresai

Eilės numeris	**Naujienų šaltinio pavadinimas**	**Šaltinio RSS adresas**
1	Delfi	http://www.delfi.lt/rss/feeds
2	Lrytas	http://www.lrytas.lt/rss/
3	15min	http://www.15min.lt/rss
4	Elektronika	http://www.elektronika.lt/rss/visas/
5	Technologijos	http://feeds2.feedburner.com/technologijos-visos-publikacijos
6	Alkas	http://alkas.lt/feed/
7	Verslo žinios	http://vz.lt/RSS.aspx
8	Pinigų karta	http://www.pinigukarta.lt/feed
9	Veidas	http://www.veidas.lt/feed
10	Politika	http://politika.lt/feed/
11	Litas.lt	http://www.litas.lt/feed/
12	Penki.lt	http://www.penki.lt/lt

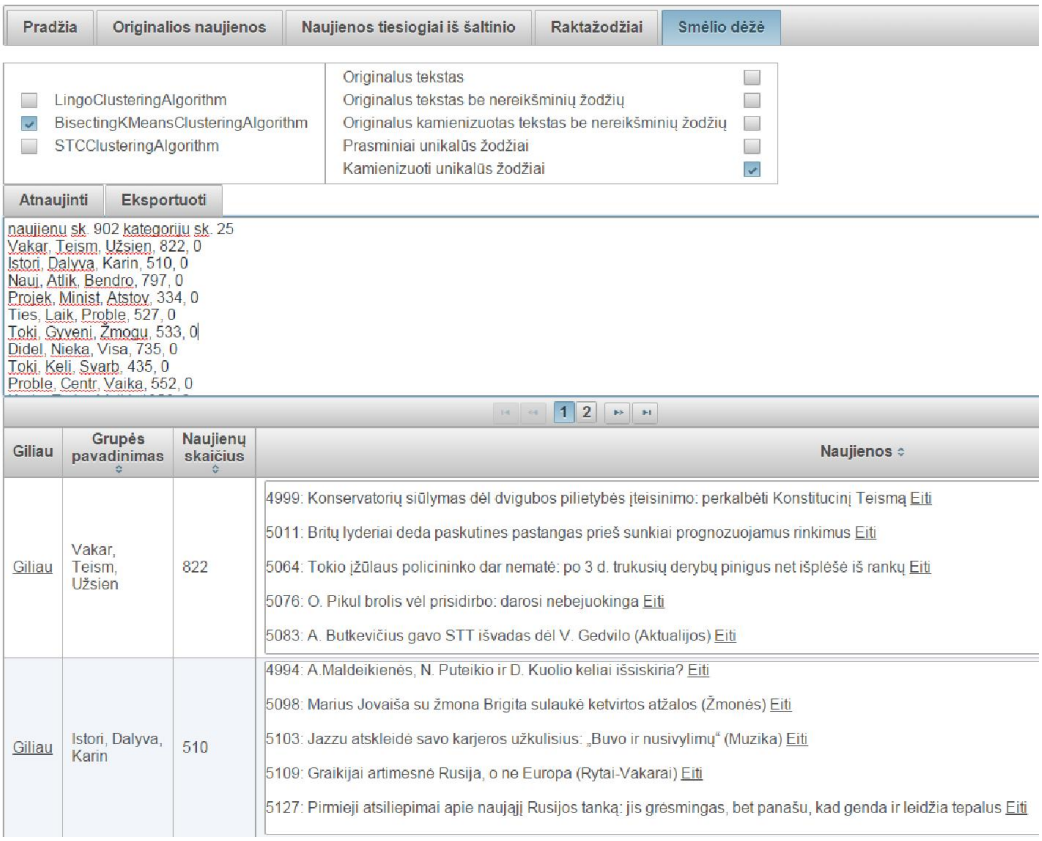
Sistemos vaizdai



12 pav. Pradinis sistemos langas



13 pav. Naujienų ir jų kategorijų, kurias pateikia šaltiniai, langas



14 pav. Naujienų klasifikavimo langas

Eksperimento metu generuotų naujienų grupių pavadinimų ir priskirtų naujienų skaičiaus rezultatai

9 lentelė Naujienų grupės gautos naudojantis „Lingo“ algoritmą

T1	**T2**	**T3**	**T4**	**T5**
naujienu sk. 902 kategoriju sk. 16 Gal, Nuotr, Jūsų, 803 Tik, Kaip, Savo, 901 Jie, Kurie, Nes, 794 Šiuo Metu, 294 Lietuvos, 501 Bus, 479 Prie, 455 Net, 425 Būtų, 391 Vis, 380 Vienas, 378 Todėl, 371 Dabar, 368 Kuris, 363 Kur, 361 Kita, 1	naujienu sk. 902 kategoriju sk. 16 Naudoti, Autoriai, Verslas, 439 Agentūros, Informavimo Priemonėse Tinklalapiuose Raštiško, Draudžiama, 88 Arenų Dalinkis Įžvalgomis Komentarais, 48 Patys, 176 Rūšis Standartinė Publikacija, 46 Kuriuos, 172 Nieko, 170 Tokia, 169 Kurio, 162 Visus, 159 Kitas, 154 Visiškai, 152 Atveju, 151 Daryti, 146 Žinoma, 143 Kita, 93	naujienu sk. 902 kategoriju sk. 16 Naudo, Versl, Autori, 691 Informav Priemon, Priemon Tinklalapiu Raštiš, Raštiš Sutiki Draudži, 61 Toki, 407 Kuria, 332 Keli, 318 Didel, 306 Galimy, 300 Nauj, 293 Ties, 253 Gyveni, 250 Klausi, 244 Visa, 243 Koki, 242 Pradė, 242 Svarb, 236 Kita, 11	naujienu sk. 902 kategoriju sk. 16 Sąjungos, 49 Patys, 47 Žmogaus, 47 Žmogus, 47 Manau, 46 Sportas, Gyvai, 27 Žinoma, 46 Taigi, 45 Vakarų, 45 Nieko, 44 Gyventojų, 43 Užsienis, 43 Gyvenimo, 42 Amžiaus, 41 Visiškai, 41 Kita, 497	naujienu sk. 902 kategoriju sk. 16 Toki, 181 Politi, 136 Nauj, 120 Galimy, 118 Gyveni, 117 Sutiki, Raštiš, 61 Keli, 112 Kuria, 112 Klausi, 109 Didel, 101 Valsty, 99 Tyrim, 97 Istori, 94 Vakar, 87 Nieka, 86 Kita, 222

10 lentelė Naujienų grupės gautos naudojantis „STC“ algoritmą

T1	T2	T3	T4	T5
naujienu sk. 902	naujienu sk. 902	naujienu sk. 902	naujienu sk. 902	naujienu sk. 902
kategoriju sk. 25	kategoriju sk. 25	kategoriju sk. 25	kategoriju sk. 25	kategoriju sk. 25
Vaikų, Jie, Nes, 52	Pergalės, Saugumo, Ukrainoje, 794	Vaika, Gyveni, Maist, 138	Gyventojų, Informacijos, Klaipėdos, 788	Vakar, Teism, Užsien, 822
Darbo, Mes, Turi, 48	Automobilio, Pinigų, Savivaldybės, 415	Įstaty, Savivaldy, Asmen, 134	Užsienis, Moteris, Komanda, 790	Istori, Dalyva, Karin, 510
Gali, Būti, Lietuvoje, 47	Maisto, Žemės, Gyventojų, 406	Vaika, Moter, Universite, 132	Kartais, Lietuvą, Bendrovės, 877	Nauj, Atlik, Bendro, 797
Lietuvos, Vilniaus, Sporto, 47	Žmogus, Žmogaus, Žinoma, 1263	Automobi, Vairuoto, Keli, 123	Sąjungos, Visiškai, Pajėgų, 848	Projek, Minist, Atstov, 334
Minlt, Val, Gegužės, 46	Sveikatos, Gyvenimo, Žmogaus, 1734	Putin, Istori, Vakar, 111	Vakarų, Ukrainoje, Prezidentas, 1448	Ties, Laik, Proble, 527
Darbo, Labai, Mokslo, 44	Žmogus, Manau, Patys, 2151	Tyrim, Telefo, Toki, 112	Patys, Atrodo, Pasakojo, 863	Toki, Gyveni, Žmogu, 533
Karo, Buvo, Bus, 44	Gyvenimo, Manau, Kalba, 1237	Karin, Putin, Techni, 109	Nieko, Visiškai, Daryti, 783	Didel, Nieka, Visa, 735
Buvo, Jis, Sakė, 43	Gyvenimo, Kalba, Žmogaus, 692	Teism, Moter, Pinig, 108	Patys, Žmogaus, Žmogus, 1125	Toki, Keli, Svarb, 435
Proc, Jav, Darbo, 43	Teismo, Teismas, Klaipėdos, 650	Sezon, Nauj, Pergal, 98	Nieko, Dažnai, Atrodo, 1293	Proble, Centr, Vaika, 552
Minlt, Val, Prieš, 40	Sveikatos, Studijų, Klaipėdos, 694	Projek, Galimy, Nauj, 96	Vakarų, Sovietų, Pergalės, 1243	Kuria, Tyrim, Veikl, 1356
Proc, Tūkst, Eurų, 40	Gyvenimo, Amžiaus, Vaikai, 1218	Veikl, Galimy, Bendro, 85	Žmogus, Taigi, Nieko, 1680	Gyveni, Gyven, Moter, 575
Lietuvos, Vilniaus, Gegužės, 38	Amžiaus, Vaikai, Patys, 1154	Vakar, Sovie, Pergal, 82	Galbūt, Saugumo, Teigimu, 1281	Tyrim, Atlik, Atvej, 506
Mūsų, Labai, Mes, 38	Kultūros, Kalba, Žinoma, 373	Politi, Valsty, Ekonomi, 142	Sąjungos, Sovietų, Prezidentas, 1493	Vakar, Visa, Koki, 427
Rusijos, Karo, Jav, 37	Šeimos, Gyvenimo, Žmogaus, 646	Sezon, Tašk, Pergal, 66	Manau, Galbūt, Saugumo, 1226	Kuria, Didel, Žinom, 400
Minlt, Val, Prieš, 36	Taigi, Žmogaus, Žinoma, 736	Politi, Tyrim, Minist, 62	Vakarų, Kalba, Saugumo, 836	Toki, Didel, Visa, 1376
Rusijos, Buvo, Mes, 36	Šeimos, Taigi, Tarnybos, 268	Vakar, Keli, Nauj, 52	Patys, Nieko, Pinigų, 2103	Visa, Koki, Žmogu, 525
Mūsų, Valstybės, Europos, 35	Gyvenimo, Pinigų, Žmogaus, 474	Politi, Istori, Klausi, 100	Skaičius, Galbūt, Saugumo, 577	Tyrim, Asmen, Atlik, 1454
Buvo, Jav, Sakė, 32	Kultūros, Gyvenimo, Taigi, 574	Gyveni, Ekonomi, Pinig, 88	Galbūt, Šiemet, Komanda, 1323	Didel, Tyrim, Žinom, 1984
Buvo, Vilniaus, Sakė, 30		Politi, Valsty, Klausi, 65	Prezidentas, Saugumo, Teigia, 1169	Tyrim, Veikl, Įstaty, 1296
Jei, Bus, Tiek, 29		Moter, Sąraš, Šaukti, 492	Manau, Žinoma, Galbūt, 1300	Tyrim, Atlik, Žinom, 962
Rusijos, Minlt, Jav, 27		Sąraš, Šaukti, Tarny, 132	Sąjungos, Užsienis, Sistemos, 397	Visa, Vaika, Atvej, 362
Darbo, Sakė, Sporto, 22		Sąraš, Šaukti, Tarny, 228		Atlik, Veikl, Atstov, 1120
Vaikų, Vilniaus, Lietuvoje, 21		Tyrim, Sąraš, Savivaldy, 42		Teism, Atlik, Veikl, 568
Proc, Minlt,				Toki, Tyrim, Veikl, 1024

11 lentelė Naujienų grupės gautos naudojantis „Bisecting K-means“ algoritmą

T1	**T2**	**T3**	**T4**	**T5**
naujienu sk. 902	naujienu sk. 902	naujienu sk. 902	naujienu sk. 902	naujienu sk. 902
kategoriju sk. 83	kategoriju sk. 88	kategoriju sk. 89	kategoriju sk. 80	kategoriju sk. 87
Turi Tiek, 171	Patys, 176	Projek Projek, 173	Sąjungos, 49	Tyrim, 97
Klasikinė Minlt, 164	Kalba Kalba, 116	Minist Minist, 155	Patys, 47	Proble, 74
Vilniaus Lietuvos, 156	Tarnybos, 83	Pinig Pinig, 153	Žmogaus, 47	Atlik, 73
Nuo Val, 151	Tokie Patys, 56	Įstaty Įstaty, 136	Žmogus, 47	Užsien, 67
Prieš Europos, 149	Žinoma Nieko, 47	Vaika Vaika, 127	Manau, 46	Visa, 67
Jav Buvo, 136	Sunku Žinoma, 43	Sveika Sveika, 120	Žinoma, 46	Moter, 65
Karo Buvo, 136	Manau Turime, 40	Keli Didel, 117	Taigi, 45	Koki Toki, 38
Tačiau Kiek Jis, 123	Vakarų Šalys, 35	Didel Klausi, 101	Vakarų, 45	Taig Toki, 27
Kaip Minlt, 118	Kitas Žmogus, 34	Koki Nauj, 97	Nieko, 44	Toki Vien, 25
Dėl Vaikų, 117	Gyventojų Skaičius, 33	Didel Politi, 94	Gyventojų, 43	Klausi Proble, 21
Prie Vilniaus, 116	Manau Tokie, 33	Laik Keli, 91	Užsienis, 43	Ekonomi Sąjung, 18
Juk Jie, 115	Patys Esame, 33	Toki Versl, 91	Gyvenimo, 42	Vakar Ekonomi, 18
Kad Mokslo, 114	Taigi Turime, 32	Gyveni Laik, 90	Amžiaus, 41	Parlame Politi, 16
Daugiau Rusijos, 111	Tokio Žmogaus, 32	Klausi Žinom, 90	Kalba, 41	Pratyb Karin, 16
Net Vilniaus, 108	Žmogaus Gyvenimą, 32	Atlik Tyrim, 89	Visiškai, 41	Kuria Visa, 15
Pagal Europos, 106	Žmogaus Teisių, 32	Klausi Ties, 89	Skaičius, 40	Putin Politi, 15
Prie Proc, 100	Sąjungos Prezidentas, 31	Padėt Toki, 89	Ukrainoje, 40	Tėva Vaika, 15
Nuotr Lietuvoje, 99	Vakarų Europoje, 30	Gyveni Politi, 85	Kartais, 39	Advoka Teism, 14
Visų Lietuvoje, 99	Sąjungos Valstybių, 29	Visa Galimy, 85	Sovietų, 39	Klausi Reikė, 14
Lietuvoje Proc, 98	Taigi Svarbu, 28	Klausi Koki, 84	Galbūt, 38	Atlik Krašt, 13
Prie Rusijos, 98	Karas Ukrainoje, 27	Nauj Žinom, 83	Informacijos, 38	Didel Priemo, 13
Prieš Kitas, 98	Karinių Pajėgų, 27	Putin Putin, 80	Lietuvą, 38	Karin Ginkl, 13
Valstybės Arba, 98	Duomenimis Gyventojų, 26	Sukur Nauj, 80	Atveju, 37	Kita Gyveni, 13
Metu Jav, 97	Teismui Teismas, 25	Valsty Koki, 79	Bendrovės, 37	Maskv Politi, 13
Pagal Darbo, 97	...	Svarb Istori, 77	Klaipėdos, 37	Didel Pradž, 12
Prieš Metus Metais, 97		Žinom Politi, 75	Prezidentas, 37	Galimy Sprend, 12
...		...	...	...

PRASMINIŲ ŽODŽIŲ ANALIZĖ LIETUVIŠKŲ NAUJIENŲ SRAUTE

Ernesta Kebelytė[1], **MantasLukoševičius**[1, 2]

[1]*Kauno technologijos universitetas, Programų inžinerijos katedra, Studentų g. 50, Kaunas,*
[2]*KTU, Biomedicininės inžinerijos institutasK. Baršausko 59Lietuva,*
ernesta.kebelyte@ktu.edu,mantas.lukosevicius@ktu.lt

Santrauka. Dėl kasdien įvairiuose internetiniuose šaltiniuose sukuriamo didelio kiekio naujienų, skaitytojams sunku išsirinkti ir lyginti pateikiamą informaciją. Atsiranda poreikis kurti algoritmus ar sistemas, kurios geba analizuoti ir apdoroti didelius tekstinius informacijos srautus.Šiame darbe pateikiama lietuviškų naujienų tekstų analizė siekiant sukurti „protingesnę“ naujienų paiešką. Darbo akcentas yra realizuoti ir ištirti teksto mažinimo algoritmus, paliekant prasminius žodžius. Teksto kiekio mažinimui pasitelkiama nereikšminių žodžių šalinimo ir panašių žodžių jungimu naudojantis žodžių kamienizavimo algoritmais.

Raktiniai žodžiai: RSS, žodžių kamienizavimas,klasterizavimas, dažnių analizė, lietuviškos naujienos.

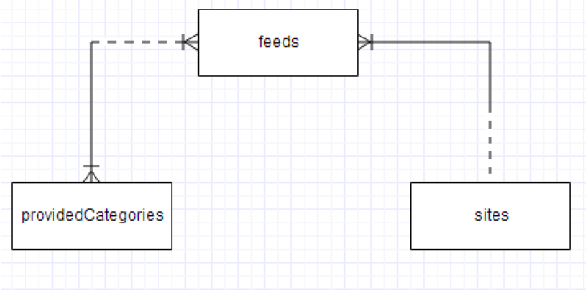
1. Įžanga

Galima lengvai pastebėti, kad internete gyvuoja daugybė informacijos platinimo portalų, kurie turinio prasme pralenkia popierinius analogus.Kasdien sukuriamas didelis informacijos srautas, kurį be įvairių programinių įrankių būtų sunku aprėpti. Kaip vienas iš palengvinančių informacijos surinkimo ir pateikimo būdų yra RSS (angl. „*Rich Site Summary*“) prenumeratos. RSS skirtas surinkti bei apdorojus pateikti naujausius įrašus iš įvairių informacijos šaltinių. Naujienų agregatoriai iškilo internetinėje erdvėje dėl šių priežasčių: saugu, praktiška [1]. Gaunama informacija yra grupuojama pagal kategorijas, tačiau čia atsiranda daugybė vietos interpretacijoms: kokios turėtų būti grupės, kaip skirstyti ir susieti skirtingą informaciją. Galiniam vartotojui svarbiausia yra gauti jam norimą informaciją ir aktualijas, per daug nešvaistant laiko.

Šiame darbe reikalinga informacija surenkama pasitelkiant RSS naujausios informacijos, pateikiamos XML formatu, surinkimo iš internetinių portalų technologija. Naujienų įrašai pagal pareikalavimą yra parsisiunčiami iš žinomų šaltinių. Dėl didelio informacijos kiekio buvo pasirinkta sukurti algoritmus, kurie „mokintųsi“ atskirti prasmingą informaciją ir susieti panašią reikšmę turinčius žodžius. Šio darbo tikslas yra mažinti lietuviškų naujienų tekstus bei analizuoti kaip tai įtakoja paiešką pagal raktinius žodžius.

2. Tekstų surinkimas

Informacijos surinkimui naudojama 33 RSS šaliniai, tačiau bendras naujienų portalų skaičius yra 13 (delfi.lt, elektronika.lt, technologijos.lt ir kt.), nes kai kurių portalų RSS srautai skirstomi pagal tematiką (Mokslas, Verslas). Pradiniame naujienų surinkimo etape iš RSS srauto surenkama tokia informacija: naujienos pavadinimas, nuorodą į originalų šaltinį, naujienos trumpas aprašymas, publikavimo data [2] bei, pilnas naujienos tekstas. Surinkta informacija apdorojama ir saugoma duomenų struktūroje, kuri pavaizduota 2 paveikslėlyje. Šiam tyrimui pasirinktas konkretaus laikotarpio naujienų srautas: nuo 2015-03-04 iki 2015-03-07 dienos. Šiame laikotarpyje iš viso gautos 1591 naujienos, tačiau svarbu paminėti, kad tyrimui naudotos naujienos, kuriose tekstai susideda daugiau nei iš 300 žodžių. Palyginimui, naujienų tekstuose žodžių skaičius svyruoja nuo kelių tūkstančių iki kelių dešimčių, kadangi kai kurios naujienos turi tik trumpą komentarą ir kartu pateikiamą video reportažą.



1 pav. Esybių ryšių diagrama

3. Prasminių žodžių išskyrimas

Kuriant autonomiškas sistemas, vienas iš pasiruošimo darbų yra tinkamai pateikti duomenų rinkinius. Naujienų straipsniai ir tematikos yra lengvai suprantamos žmonėms, tačiau tam, kad mašina galėtų sukurti semantinius ryšius tarp duomenų rinkinių, klasifikuoti ar sisteminti informaciją, pradiniai duomenys turi būti apdoroti. Literatūroje [3] pateikiami pagrindiniai pradinių duomenų apdorojimo metodai:

4. Pašalinti bendrieji žodžiai (jungtukai, įvardžiai ir t.t.);
5. Sukuriamas žodynas ar komponentas susiejantis panašių reikšmių žodžius (pvz. tą patį žodžio kamieną turintys žodžiai reprezentuoja bendrą prasmę).

Surinktų lietuviškų naujienų tekstų analizė apima dažniausių žodžių bei žodžių dažniausių galūnių rinkimą. Naujienų tekstų apdorojimas atliekamas tokia tvarka: atliekamas naujienų teksto mažinimas atmetant nereikšminius žodžius, naujienų tekstuose esantiems likusiems žodžiams atliekama galūnių analizė ir žodžių kamienizavimas.

Kitas svarbus aspektas analizuojant naujienų tekstus skaičiuoti, kurie žodžiai yra svarbiausi konkrečioje naujienoje. Sistema naujienos tematiką gali suprasti pagal joje rastus pasikartojančius reikšminius žodžius. Jei naujienoje konkretaus žodžio pasikartojimo dažnis yra ženkliai didesnis už kitų žodžių dažnį, tuomet toks žodis yra reikšminis ir gali nurodyti naujienos kategoriją ar tematiką.

3.1. Dažniausiai pasikartojančių kalbos žodžių išmetimas

Informacijos srauto mažinimui sukurtas dažniausių žodžių atmetimo algoritmas. Analizuojant pradinius naujienų tekstus renkama informacija apie visus panaudotus žodžius tekstuose. Rezultate gaunamas žodynas, kuriame kiekvienas žodis turi pasikartojimo skaičių - dažnį visame naujienų sraute. Pagal turimą informaciją išrenkami dažniausi žodžiai naudojami naujienų tekstuose (2 pav.) Darant prielaidą, kad 1/5-ąją naujienų teksto dalį sudaro dažniausi žodžiai galima išvesti formuluotę, pagal kurią sistema atrenka dažniausius žodžius iš sukurto žodyno. Turint sąrašą dažniausių žodžių, kurie yra nereikšminiai, apdorojami visi naujienų tekstai atmetant nereikšminius žodžius.

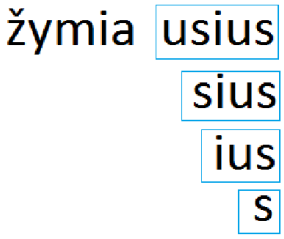
ir	10645	savo	1343		
kad	3741	tik	1271		
į	2752	bet	1235		
su	2258	apie	1218		
iš	2230	taip	1199		
tai	2097	nuo	1049		
yra	2018	tačiau		1027	
o	1717	per	956		
kaip	1603	lietuvos		942	
buvo	1573	dar	941		
metu	1515	kai	934		
ar	1435	jau	897		
ne	1352	jis	894		

2 pav. Dažniausi žodžiai

3.2. Žodžių kamienizavimas

Naujienų tekstuose pilna pasikartojančių bei panašių žodžių, kurie rodo tą pačią informaciją. Kadangi formaliai žiūrint, žodžiai kaip „žymiausius“, „žymiausių“, „žymiausias“, „žymiausiais“, suprantami kaip pavieniai ir nesusiję, reikalingas algoritmas gebantis sujungti tą pačią prasmę turinčius žodžius.

Informacijos kiekio mažinimui pasitelkiamas žodžių galūnių atskyrimas, atliekamas pasinaudojant jau apdorotais naujienų tekstais, kuriuose yra tik prasminiai žodžiai. Toliau atliekama žodžių pabaigų analizė: kiekvienas naujienos teksto žodis dalinamas į dvi dalis. Antrosios žodžio dalies raidžių (3 pav.) deriniams skaičiuojami pasikartojimo dažniai visame naujienų žodžių sraute. Analizės pabaigoje gaunamas dažniausių galūnių ir jų dažnių sąrašas (5 pav.).



3 pav. Žodžio galūnės analizė

"s";40032	"os";9426
"i";24621	"ai";8276
"o";17122	"is";6369
"a";12697	"ti";5832
"e";10758	"us";5566
"as";9613	"jos";2248

4 pav. Dažniausių galūnių sąrašas

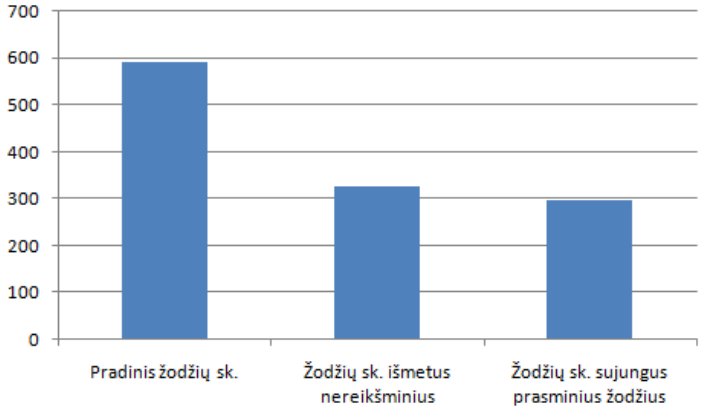
Surinkus duomenis apie dažniausias galūnes, atliekama konkrečių naujienų tekstų analizė, kurioje analizuojamas kiekvienas teksto žodis: ieškoma didžiausia galūnė, kuri atitiktų nagrinėjamo žodžio pabaigą. Galūnė negali būti ilgesnė nei pusė žodžio. Rasta galūnė atmetama nuo nagrinėjamo žodžio pabaigos.

3.3. Skirtingų žodžių mažinimo rezultatai

Bendras žodynas, kuris gautas analizuojant visus naujienų tekstus, sudaro apie 83 tūkstančiai unikalių žodžių. Atlikus žodžių kamienizavimą gautas apie 48 tūkstančių žodžių kamienų žodynas. Galima pridurti kad ne visi naujienose rasti žodžiai yra lietuviški, galima aptikti žargonizmų, tačiau tai labiau priklauso nuo naujienų autoriaus ir tematikos.

5 paveikslėlyje pateikiami naujienų tekstų žodžių vidurkiai: prieš analizę, po nereikšminių žodžių šalinimo bei prasminių žodžių sujungimo atliekant kamienizavimą. Primename, kad analizei pasirinkti naujienų tekstai, kuriuose pradinis žodžių skaičius yra didesnis negu 300 žodžių. Svarbu paminėti, kad po nereikšminių žodžių šalinimo saugomi tik unikalūs žodžiai kartu su jų pasikartojimo skaičiumi tekste, taip išvengiama mums naudingos informacijos dubliavimosi.

Matomas ženklus žodžių sumažėjimas po dažniausių žodžių atmetimo procedūros: vidutiniškai naujienos tekstai sumažėjo 45 procentais, o atlikus žodžių kamienizavimą informacija apie naujieną nuo pradinio kiekio sumažėjo 50 procentų. Tokie rezultatai gali ženkliai pagreitinti naujienų paiešką pagal raktinius žodžius, taip pat gali pagerinti naujienų paiešką. Grafike (5 pav.) matomas nedidelis žodžių skaičiaus sumažėjimas tarp antro ir trečio stulpelio dėl to, kad saugomi unikalūs naujieną sudarantys žodžiai kartu su jų pasikartojimo dažniu. Tokio rezultato buvo tikėtasi, kadangi naujienų tekstų didžiąją dalį sudaro mažai besikartojantys žodžiai, o naujienos tematiką reprezentuojantys žodžiai dažnai sudaro tik 5 procentus teksto.



5 pav. Naujienų tekstų mažinimo rezultatai

4. Straipsnių paieška pagal prasminius žodžius

Atlikome naujienų paieškų eksperimentus pagal reikšminius žodžius. Pirmu eksperimentu atliekame ir lyginame paiešką naudojantis dažnai naujienų sraute randamais raktažodžiais ir jų kamienizuota forma. Antrame eksperimente lyginame paiešką pasirinkus konkrečius raktažodžius su paieška naudojantis kamienizuotais raktažodžiais.

Pirmame eksperimente iš bendro žodyno, kurį sudaro visi žodžiai naudojami naujienų tekstuose išsirinkę kelis raktinius žodžius, kurie pasirodė naujienų sraute dažniausiai, galime sukurti palyginimo lentelę. Lentelės ašyse atvaizduojami raktiniai žodžiai bei naujienos, kuriuose šie raktiniai žodžiai pasikartojo daugiausiai. Eksperimentui pasirenkama raktiniai žodžiai bei jų kamienizuota formos nurodytos 1 lentelėje. Atrenkamos kelių dienų naujienos pagal pasirinktus raktinius žodžius. Gauti rezultatai pateikiami 1 lentelėje.

Lentelė Nr.1 Paieškų pagal raktinį žodį rezultatai

	ukrainoje (ukrain)	rusija (rusij)	karių (kari)	vilniuje (vilniu)	kaziuko (kaziu)
R. Murmokaitė: Rusija sukėlė karą Ukrainoje ir...	X(X)	X(X)			
M. Saakašvilis: Baltijos šalys karo atveju neturi tokio...	X(X)	X(X)	X(X)		
Košmaras Ukrainoje: Debalcevėje rasti 500 civilių kūnų	X(X)	X(X)			
Rusijos gynybos viceministras maivėsi kalbėdamas apie...	X(X)	X(X)	X(X)		
ESBO raginama dvigubinti savo gretas Rytų Ukrainoje	X(X)				
NATO pareigūnai skaičiuoja Ukrainoje žuvusius rusų...	X(X)		X(X)		
NATO pareigūnas: Rusijos gyvosios jėgos nuostoliai...	X(X)	**(X)**	X(X)		
Vilniaus meras – prieš privalomą priešmokyklinį ugdymą				X(X)	
Vilniuje įsišėlo Kaziuko mugė: kiek teks pakloti				X(X)	X(X)
Kaziuko mugėje riedės XIX a. karieta				X(X)	X(X)
Vasarį prognozuoja mažesnes sąskaitas už šilumą nei...				X(X)	
Kaziuko mugėje riedės XIX a. karieta (mugės programa)				X(X)	X(X)

Iš gautų rezultatų matoma, kaip pagal raktinius žodžius atrinktos naujienos skirstosi į atskirus klasterius. Todėl galima daryti prielaidą, kad pasinaudojus sudarytu žodynu ir turima informaciją apie naujienų dažniausius žodžius galima klasterizuoti lietuviškų naujienų srautą. Šiuo atveju galima išskirti konkrečius naujienų klasterius: [unkrainoje, rusija, karių], [vilniuje, kaziuko]. Taip pat išryškėja kamienizavimo privalumai: panaudojus kamienizuotą raktažodį „rusij“ atrasta naujiena, kuri nebūtų iškilusi naudojantis raktažodžiu „rusija“. Tokiu būdu galima sukurti tikslesnius naujienų klasterius.

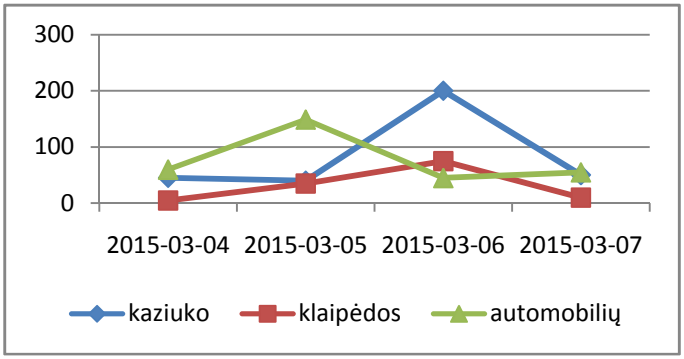
Tikslesni paieškų rezultatai pateikiami 2 lentelėje. Iš gautų paieškų rezultatų matomas naujienų padidėjimas jei paieška atliekama pagal kamienizuotą raktažodį. Naujienų, kuriuose panašią prasmę turintys žodžiai sujungiami, galutinėje paieškoje randama 22 procentais daugiau. Tokio rezultato buvo tikimasi, kadangi tik galūnėmis tesiskiriantys žodžiai dažnai vartojami naujienų tekstuose.

Lentelė Nr. 2 Paieškos pagal pilnus raktažodžius ir jų kamienus rezultatai

Raktažodis	**Naujienų skaičius**	**Raktažodžio kamienizuota dalis**	**Naujienų skaičius**
ukrainoje	72	ukrain	83
rusija	63	rusij	98
karių	26	kari	28
vilniuje	60	vilniu	60
kaziuko	14	kaziuk	18

Kita svarbi įžvalga, kurią galima gauti pasinaudojant surinkta informacija yra naujienų temų aktualumo analizė. Įvairūs įvykiai ar renginiai iškyla tik konkrečiu laikotarpiu ir naujienų temos yra aktualios tik savaitę ar kelias dienas, taip pat galima daryti įvairią analizę pagal tai, koks konkrečių raktažodžių dažnių pasiskirstymas laike. Keli raktažodžių dažnių pasiskirstymai laike pateikti 6 paveikslėlyje.

Pateiktoje diagramoje (6 pav.) galima pamatyti kaip iškyla raktažodis „kaziuko“, toks iškylimas įvyko dėl Vilniuje rengiamos tautodailininkų mugės. Raktažodžio „automobilių“ iškilimą galėjo lemti Ženevos automobilių paroda, kiti raktažodžiai iškyla dėl įvairių renginių, nelaimingų atsitikimų statistikos ar kita. Informacijos pasiskirstymas laike labai patogus norint sekti konkrečius įvykius, kurių informacija lietuviškuose naujienų šaltiniuose gali keistis kelių dienų bėgyje.



6 pav. Prasminių žodžių dažnių pasiskirstymas laike

5. Išvados ir tolesni darbai

Šiame darbe buvo pateiktas sprendimas, kaip galima sumažinti tekstą dideliame naujienų sraute. Aptarti du pagrindiniai tekstų mažinimo metodai: nereikšminių žodžių šalinimas bei žodžių kamienizavimas. Atliktame eksperimente pastebėtas iki 50 procentų informacijos sumažinimas neprarandant tekstuose pateikiamų reikšminių žodžių. Įrodytą, jog panašių žodžių sujungimas atliekant kamienizavimą gali padėti sukurti „protingesnę“ naujienų paiešką.

Tolesniame darbe žadame: sukurti pilnai automatinį naujienų klasterizavimą pagal kelis susijusius reikšminius žodžius, pateikti aktualiausias naujienas pasitelkiant reikšminių žodžių dažnių kaitą laike. Taip pat šis darbas atveria plačias perspektyvas pateikti surinktą informaciją įvairiais pjūviais.

Literatūros sąrašas

[1] **Simec A., Carapina M., Duk S.**RSS as medium for information and communicationtechnology.*MIPRO, 2011 Proceedings of the 34th International Convention* , pp.1593,1596, 23-27 Gegužė 2011.

[2] **Han Y. G., Lee S. H., Kim J. H., KimY.**A New Aggregation Policy for RSS Services. *Proceedings of the 2008 international workshop on Context enabled source and service selection, integration and adaptation: organized with the 17th International World Wide Web Conference*, Article No. 2 , 2008.

[3] **Steinbach, M.; Karypis, G., Kumar, V.**A Comparison of Document Clustering Techniques. *Technical report, University of Minnesota*, 2000.

Keywords analysis of Lithuanian news stream

Abstract -On a daily basis a large amount of newsis generated in various online websites. As a result, it is difficult for users to choose news that are important and compare them with other information. There is a growing need to develop algorithms or systems that are able to analyze and process large texts and information flows. This paper presents analysis of Lithuanian news texts in order to create "smart" news search. This study focuses on text reduction algorithms and its results. Reducing the volume of text in Lithuanians news is executed through the elimination of non-essential words and aggregation ofsimilar words using word stemming algorithms.